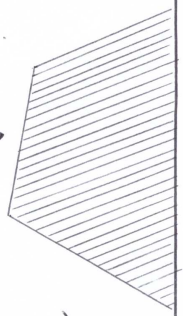
SL = SUPPORT VECTOR MACHINE

I- ELEMENTS OF CONVEX OPTIMIZATION

I.1. Convex Sets & Convex Functions.

→ A convex set C contains line segments between two points in the set :

$\forall u, v \in C, \quad 0 \leq \lambda \leq 1, \quad \lambda u + (1-\lambda) v \in C$.

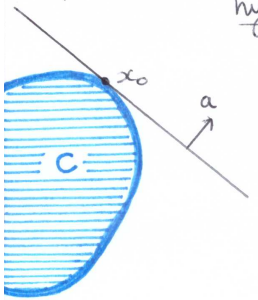


• Ex: A polyhedron = intersection of halfspaces.

(Operations that preserve convexity are: intersection, image of a convex set under affine transformations $x \mapsto Ax + b$; linear fractional function $f : x \mapsto \frac{Ax+b}{c^t x + d}$ ← >0)

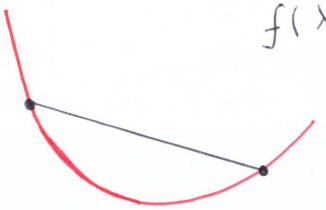
Result: If C is convex, then there exists a supporting hyperplane at every boundary point of C.

↖ at x_0 : $\{ x \mid a^t x = a^t x_0 \}$, $a \neq 0$
s.t. $a^t x \leq a^t x_0 \quad \forall x \in C$



→ A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if its domain is a convex set, and $\forall x, y \in \text{dom} f$, $\forall 0 \leq \lambda \leq 1$,

$f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$



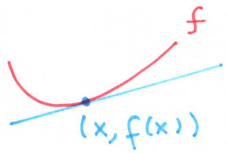
Ex: $|x|^p$ on $\mathbb{R}$,

$\|x\|_\infty = \max_k |x_k|$

$\|x\|_p = \left(\sum |x_i|^p \right)^{1/p}$, $p \geq 1$

• First order condition (f differentiable)

$f(y) \geq f(x) + \nabla f(x)^t (y - x) \quad \forall x, y$



• Second order condition (twice diff).

f is (strictly) convex $\iff \nabla^2 f(x) \succcurlyeq 0 \quad \forall x$
$(\succ)$

• Ex: (i) Quadratic functions $f(x) = \frac{1}{2} x^t P x + q^t x + r$

$\nabla f(x) = Px + q$

$\nabla^2 f(x) = P$

f is convex iff $P \succcurlyeq 0$.

(ii) Least-Squares criterion $f(x) = \| Ax - b \|_2^2$

$\nabla f(x) = 2A^t (Ax - b)$

$\nabla^2 f(x) = 2A^t A$

$\Rightarrow$ convex $\forall A$

(iii) Log-Sum exponential $f(x) = \log \left(\sum_{k=1}^{d} e^{x_k} \right)$.

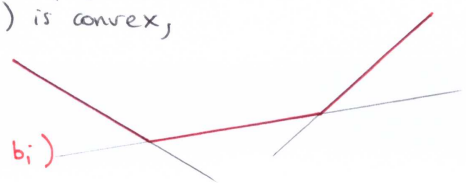
After calculations,

$\nabla^2 f(x) = \frac{1}{1^t z} \text{diag}\, z - \frac{1}{(1^t z)^2} z z^t, \quad z_k = e^{x_k}$

We must check that $\forall u, \quad u^t \nabla^2 f(x)\, u \geq 0$.

$u^t \nabla^2 f(x)\, u = \frac{\left(\sum z_k v_k^2 \right)\left(\sum z_k \right) - \left(\sum v_k z_k \right)^2}{\left(\sum z_k \right)^2} \geq 0$ (↑ CS)

Operations that preserve convexity = non-negative weighted sum, composition with an affine function : $f(Ax+b)$ is convex if f is convex, pointwise maximum: if $f_1, \dots, f_k$ are convex, then $f(x) := \max (f_1(x), \dots, f_k(x))$ is convex, pointwise supremum.



$f(x) = \max_i (a_i^t x + b_i)$

proof for pointwise supremum: if $f(x,y)$ is convex in x for all $y \in \mathcal{Y}$, then $g(x) := \sup_{y \in \mathcal{Y}} f(x,y)$ is convex:

Let x_1, x_2, $0 \leq \lambda \leq 1$,

$$g(\lambda x_1 + (1-\lambda) x_2) = \sup_{y \in \mathcal{Y}} f(\lambda x_1 + (1-\lambda) x_2, y)$$

$$\leq \sup_{y \in \mathcal{Y}} \{\lambda f(x_1, y) + (1-\lambda) f(x_2, y)\}$$

$$\leq \lambda \sup_{y \in \mathcal{Y}} f(x_1, y) + (1-\lambda) \sup_{y \in \mathcal{Y}} f(x_2, y)$$

$$= \lambda g(x_1) + (1-\lambda) g(x_2).$$

since $\sup(x_n + y_n) \leq \sup x_n + \sup y_n$, as $\forall n$ $x_n + y_n \leq \sup x_n + \sup y_n$ (independent of n).

[Ref] Boyd & Vandenberghe (04)

I.2. Convex Optimization Problems.

An optimization problem (not necessarily convex) in standard form:

minimize	$f_0(x)$	← $x \in \mathbb{R}^d$ = decision variable
subject to	$f_i(x) \leq 0$ $\quad i = 1, \dots, m$	
	$h_i(x) = 0$ $\quad i = 1, \dots, p$	

$f_0 : \mathbb{R}^d \to \mathbb{R}$ = objective / goodness-of-fit / cost

$f_i : \mathbb{R}^d \to \mathbb{R}$ = inequality constraints

$h_i : \mathbb{R}^d \to \mathbb{R}$ = equality constraints

) prior information / parameter limits.

↳ A point x is FEASIBLE if $x \in \text{dom } f_0$ & satisfies the constraints. The set of all feasible points is called the feasible set.

↳ The optimal value is $p^* := \inf \{ f_0(x) \mid f_i(x) \leq 0 \ \& \ h_i(x) = 0 \}$

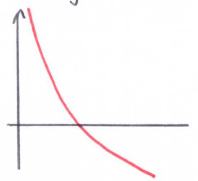
$p^* = +\infty$ if the problem is infeasible, and $= -\infty$ if unbounded below.

↳ x^* is OPTIMAL if $f_0(x^*) = p^*$.

• Ex $f_0(x) = 1/x$, $\text{dom } f_0 = \mathbb{R}_+$

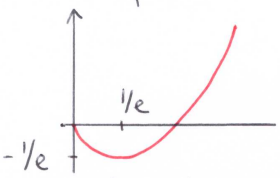
$p^* = 0$, no optimal point

$f_0(x) = -\log x$, $\text{dom } f_0 = \mathbb{R}_+$



$p^* = -\infty$

$f_0(x) = x \log x$, $\text{dom } f_0 = \mathbb{R}_+$



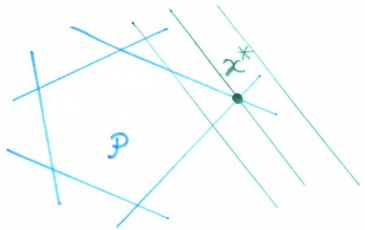
$p^* = -1/e$, $x^* = 1/e$.

• A convex optimization problem in standard form:

minimize	$f_0(x)$	↺ f_0, f_i convex
subject to	$f_i(x) \leq 0$ $\quad i = 1, \dots, m$	
	$a_i^t x = b_i$ $\quad i = 1, \dots, p$	← affine equality constraints.

• Ex: Linear Programming (LP):

minimize	$c^t x + d$	
subject to	$G x \leq h$	
	$A x = b$	



A few facts about LP:

x If a LP problem has exactly one optimal solution, then this solution must be an extreme point of the feasible region.

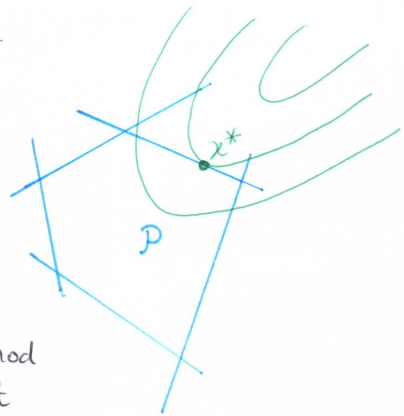
x If a LP problem has more than one optimal solution, it must have infinitely many optimal solutions. Furthermore, the set of optimal solutions is convex

x Two common approaches: Simplex Method (Dantzig, 1947): starts with an extreme point, & moves to another extreme point with an improved objective value (greedy algorithm);

Interior Point Method (Karmardar (1984)) : moves from the relative interior of the region or faces, towards the optimal solution.

- Ex : Quadratic Programming (QP).

minimize $\frac{1}{2} x^t P x + q^t x + r$
subject to $Gx \preceq h$
$Ax = b$



P = positive-semi-definite matrix

Available Algorithms = Ellipsoid Method
Interior Point
Extensions of Simplex Algorithm ... /...

I.3. Duality.

Recall the problem in standard form (real world)

minimize $f_0(x)$, $x \in \mathbb{R}^d$
subject to $f_i(x) \leq 0$, $1 \leq i \leq m$
$h_i(x) = 0$, $1 \leq i \leq p$

Put $\mathcal{D} = \bigcap_{i=0}^{m} \text{dom } f_i \cap \bigcap_{i=1}^{p} \text{dom } h_i$

Optimal value = p^*

aka the PRIMAL PROBLEM

→ We introduce the Lagrangian function $\mathcal{L}$ = linear combination of the objective & the constraints :

$$\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \longrightarrow \mathbb{R}$$
$$(x, \lambda, \nu) \longmapsto f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x).$$

↑ "Smoothes out" the original problem : if the candidate x is infeasible, at least one of the constraints are violated, which incurs an infinite cost : the original problem can be rewritten :

$$f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) + \sum_{i=1}^{p} I_0(h_i(x))$$

$$I_-(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ +\infty & \text{if } x > 0 \end{cases} \qquad I_0(x) = \begin{cases} 0 & \text{if } x = 0 \\ +\infty & \text{otherwise.} \end{cases}$$

↳ The Lagrangian avoids to jump from a finite value to an infinite one, by considering a weighted sum of the objective & the constraints.

→ The Lagrange dual function is $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$
$(\lambda, \nu) \mapsto \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu)$.

↑ "The optimal cost given λ and ν".

For each x, $\mathcal{L}(x, \lambda, \nu)$ is affine in (λ, ν), and thus both concave & convex in (λ, ν). The pointwise infimum of a concave function is concave $\Rightarrow$ g is concave (even if the original problem is not a convex optimization problem.

Lower Bound Property = If $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^*$

proof : If u is feasible, and $\lambda \succeq 0$, then

$$f_0(u) \geq \mathcal{L}(u, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \nu) = g(\lambda, \nu)$$

↑ since $f_i(u) \leq 0$, $h_i(u) = 0$

Minimizing over all feasible u gives $p^* \geq g(\lambda, \nu)$. ∎

- Ex : Least-Squares Solution

minimize $x^t x$
subject to $Ax = b$

$$\mathcal{L}(x, \nu) = x^t x + \nu^t (Ax - b)$$
$$\in \mathbb{R}^d \qquad \in \mathbb{R}^p$$

$\nabla_x \mathcal{L}(x, v) = 2x + A^t v = 0 \Rightarrow x = -\frac{1}{2} A^t v$

Plug this value of x back into $\mathcal{L}$ to get the expression of the Lagrange dual function:

$g(v) = \mathcal{L}\left(-\frac{1}{2} A^t v, v\right) = -\frac{1}{4} v^t A A^t v - b^t v$

The lower bound property gives $p^* \geq -\frac{1}{4} v^t A A^t v - b^t v \quad \forall v \in \mathbb{R}^p$

The DUAL problem is

maximize	$g(\lambda, v)$
subject to	$\lambda \geq 0$

(theoretical world)

↑ Find the best possible lower bound on p^*.

The dual problem is a convex optimization problem. We denote its optimal value d^*.
(λ, v) are dual feasible if $\lambda \geq 0$, $(\lambda, v) \in \text{dom } g$.

x Weak duality $d^* \leq p^*$ holds, always.

↳ If the primal problem is unbounded below, $p^* = -\infty$, we must have $d^* = -\infty$; i.e. the Lagrange dual problem is infeasible.

↳ If the dual problem is unbounded above, $d^* = +\infty$, we must have $p^* = +\infty$; i.e. the primal problem is infeasible.

Weak duality holds even when p^* and q^* are infinite.
The difference $p^* - d^*$ is called the duality gap.

x Strong duality $d^* = p^*$ does not hold in general.
However, strong duality usually holds for convex problems.
Conditions that guarantee strong duality in convex problems are called CONSTRAINT QUALIFICATIONS.

• Slater's constraint qualifications.

Strong duality holds for a convex problem

minimize	$f_0(x)$
subject to	$f_i(x) \leq 0$
	$a_i^t x = b_i$

if it is strictly feasible; i.e. if the exists at least one feasible point $x \in \text{int } \mathcal{D}$ that satisfies $f_i(x) < 0$, and $a_i^t x = b_i$.

Moreover, if the inequality constraints are affine, then $f_i(x) \leq 0$ is enough.

↑ If $d^* > -\infty$, Slater also guarantees that the dual optimum is attained: there exists a dual feasible point (λ^*, v^*) such that $g(\lambda^*, v^*) = p^*$ (cnvx problem).

• Ex: Back to the LS problem (page 6).

minimize	$x^t x$
subject to	$Ax = b$

p^*

maximize $-\frac{1}{4} v^t A A^t v - b^t v$

q^*

Slater's conditions hold if the primal problem is feasible: if $b \in \mathcal{C}(A)$, then $p^* = d^*$.

I.4. Optimality Conditions.

Assume that strong duality holds, and denote $x^* =$ primal optimal, $(\lambda^*, v^*) =$ dual optimal. (primal + dual optimal are attained).

$$f_0(x^*) = g(\lambda^*, v^*) = \inf_x \left\{ f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p v_i^* h_i(x) \right\}$$

$$\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p v_i^* h_i(x^*)$$

$\lambda_i^* f_i(x^*) \leq 0$; $h_i(x^*) = 0$ (x is feasible)

$$\leq f_0(x^*)$$

$\Rightarrow$ The two inequalities must hold with equality; that is

↘ x^* minimizes $\mathcal{L}(x, \lambda, \nu)$

↘ $\lambda_i^* f_i(x^*) = 0 \quad i = 1, \dots, m$

This condition is known as COMPLEMENTARY SLACKNESS.

If $\lambda_i^* > 0$, then $f_i(x^*) = 0$

while if $f_i(x^*) < 0$, then $\lambda_i^* = 0$

The i-th optimal Lagrange multiplier is zero unless the i-th constraint is active.

Summarizing, we obtain the

- KARUSH - KUHN - TUCKER (KKT) conditions.

① Primal Constraints $\quad f_i(x) \leq 0 \quad 1 \leq i \leq m$
$\quad h_i(x) = 0 \quad 1 \leq i \leq p$

② Dual Constraints $\quad \lambda \geq 0$

③ Complementary Slackness $\quad \lambda_i f_i(x) = 0 \quad 1 \leq i \leq m$

④ Gradient of Lagrangian w.r.t. x vanishes

$$\nabla_x \mathcal{L} = \nabla_x f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$$

↖ Thus, if strong duality holds, and (x^*, λ^*, ν^*) are optimal, they must satisfy the KKT conditions.
For a convex optimization problem, the converse holds as well.

For a convex optimization problem,

(x^*, λ^*, ν^*) satisfy the KKT conditions $\quad \Leftrightarrow \quad$ x^* and (λ^*, ν^*) satisfy $f_0(x^*) = g(\lambda^*, \nu^*)$ [zero duality gap]

↳ Indeed, suppose that (x^*, λ^*, ν^*) satisfy the KKT conditions.

Then x^* is feasible (from ①), and

$$\mathcal{L}(x, \lambda^*, \nu^*) = f_0(x) + \underbrace{\sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x)}_{=0, \text{ from } ①}$$

$$= f_0(x) + \sum_{i=1}^{m} \underbrace{\lambda_i^*}_{\text{all } \geq 0, \text{ from } ②} f_i(x)$$

= convex function of x
(non-negative weighted linear combination of convex functions).

Since $\mathcal{L}$ is convex, condition ④ ensures that x^* is the minimizer, and

$$g(\lambda^*, \nu^*) = \mathcal{L}(x^*, \lambda^*, \nu^*)$$
$$= f_0(x^*) + \sum_{i=1}^{m} \underbrace{\lambda_i^* f_i(x^*)}_{=0 \text{ from } ③} + \sum_{i=1}^{p} \underbrace{\nu_i^* h_i(x^*)}_{=0 \text{ from } ①}$$
$$= f_0(x^*)$$

$\Rightarrow$ x^* and (λ^*, ν^*) have zero duality gap. ∎

Remarks (i) KKT conditions used to be known as KT conditions (Kuhn - Tucker), as these conditions appeared in Kuhn & Tucker original paper in 1951. Later, people found out that Karush had these conditions appearing in his unpublished Master's thesis in 1939.

(ii) If the functions are not differentiable, we work with subgradients: the 4th condition is replaced with

$$0 \in \partial f_0(x) + \sum_{i=1}^{m} \lambda_i \partial f_i(x) + \sum_{i=1}^{p} \nu_i \partial h_i(x)$$

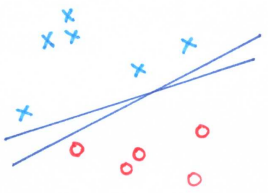
(Subgradients were introduced in the chapter SL: RR & LASSO)

↳ The equivalence statement for a convex optimization problem page 9 still holds.

II. MAXIMUM MARGIN CLASSIFIER.

II.1. The linearly separable case.

Assume in this section that the data is linearly separable.
↳ there exists infinitely many hyperplanes separating the data.



Notation: $\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, $Y_i \in \{-1, 1\}$
(binary classification problem)

A classifier constructed from $\mathcal{L}_n$ returns a new point x as $+1$ if $\beta_0 + \beta^t x \geq 0$, and as -1 otherwise.

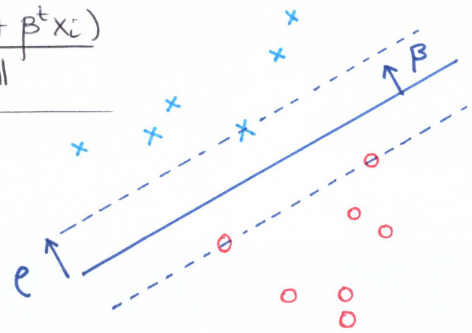
To make the solution to this problem unique, return the hyperplane that makes the largest gap (aka margin) between the two classes.

Recall that the signed distance of a point x from the hyperplane defined by $\{x \mid f(x) = \beta_0 + \beta^t x = 0\}$ is $f(x)/\|\beta\|$.

↳ distance is $\frac{y f(x)}{\|\beta\|}$; so that the margin ℓ is given by

$$\ell = \min_{1 \leq i \leq n} \frac{y_i(\beta_0 + \beta^t x_i)}{\|\beta\|}$$

Only the direction matters ⇒ restrict yourself to $\|\beta\|_2^2 = 1$



The optimization problem is

$$\begin{aligned} &\text{maximise}_{\ell} \quad \ell \\ &\text{subject to} \quad \|\beta\|_2^2 = 1 \\ &\qquad y_i(\beta_0 + \beta^t x_i) \geq \ell \quad 1 \leq i \leq n \end{aligned} \tag{0}$$

The inequality constraint is equivalent to

$$y_i\left(\frac{\beta_0}{\ell} + \left(\frac{\beta}{\ell}\right)^t x_i\right) \geq 1$$

Put $\gamma = \beta/\ell \Rightarrow \|\gamma\| = \frac{\|\beta\|}{\ell} = \frac{1}{\ell}$; $\ell = \frac{1}{\|\gamma\|}$.

Thus

maximize $\ell \Leftrightarrow$ minimize $\|\gamma\| \Leftrightarrow$ minimize $\frac{1}{2}\|\gamma\|^2$

The optimization problem (0) is equivalent to

$$\begin{aligned} &\text{minimize}_{\beta_0, \beta} \quad \frac{1}{2}\|\beta\|_2^2 \\ &\text{subject to} \quad y_i(\beta_0 + \beta^t x_i) \geq 1 \quad 1 \leq i \leq n \end{aligned} \tag{1}$$

PRIMAL PROBLEM

↑ A convex optimization problem

↳ Lagrangian is

$$\mathcal{L}(\beta_0, \beta, \lambda) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \lambda_i \{y_i(\beta_0 + \beta^t x_i) - 1\}$$

↑ The problem in standard form has inequality constraints ≤ 0.

↳ Lagrange dual function is

$$g(\lambda) = \inf_{\beta_0, \beta} \mathcal{L}(\beta_0, \beta, \lambda) \leq \frac{1}{2}\|\beta^*\|$$

↑ The primal optimum.

The data is linearly separable ⇒ strong duality holds for this convex optimization problem. Denote by λ^* the dual optimum.

⇒ $\beta_0^*, \beta^*, \lambda^*$ satisfy the KKT conditions.

- KKT conditions (linearly separable case)

① Primal constraints $1 - y_i(\beta_0 + \beta^t x_i) \leq 0 \quad i = 1, \dots, n$

② Dual constraints $\lambda \geq 0$

③ Complementary Slackness $\lambda_i \{ y_i(\beta_0 + \beta^t x_i) - 1 \} = 0 \quad i = 1, \dots, n$

④ Gradient of the Lagrangian w.r.t. β_0, β vanishes

$$\frac{\partial \mathcal{L}(\beta_0, \beta, \lambda)}{\partial \beta_0} = -\sum_{i=1}^{n} \lambda_i y_i = 0 \qquad (4.1)$$

$$\frac{\partial \mathcal{L}(\beta_0, \beta, \lambda)}{\partial \beta} = \beta - \sum_{i=1}^{n} \lambda_i y_i x_i = 0 \qquad (4.2)$$

↑ We use (4.1) and (4.2) to express β_0 & β as a function of λ, and plug these values back into $\mathcal{L}(\beta_0, \beta, \lambda)$ to obtain the expression of the Lagrange dual function $g(\lambda)$.

Note that $\beta(\lambda) = \sum_{i=1}^{n} \lambda_i y_i x_i$ (4.2), while β_0 is left unspecified.

$$\mathcal{L}(\beta_0, \beta(\lambda), \lambda) = \frac{1}{2} \|\beta(\lambda)\|_2^2 - \sum_{i=1}^{n} \lambda_i \{ y_i(\beta_0 + \beta(\lambda)^t x_i) - 1 \}$$

From (4.1), the term $\beta_0 \sum_i \lambda_i y_i$ vanishes

$$= \frac{1}{2} \|\beta(\lambda)\|_2^2 - \sum_{i=1}^{n} \lambda_i \{ y_i \beta(\lambda)^t x_i - 1 \}$$

$$= \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \lambda_i - \sum_{i,j} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle$$

$$= 1^t \lambda - \frac{1}{2} \lambda^t H \lambda = \text{quadratic function of } \lambda$$

$$1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n \qquad \lambda = \begin{pmatrix} \lambda_1 \\ | \\ \lambda_n \end{pmatrix} \in \mathbb{R}^n \qquad H = \left(y_i y_j \langle x_i, x_j \rangle \right)_{\substack{i=1,\dots,n \\ j=1,\dots,n}}$$

The dual optimal λ^* maximizes $1^t \lambda - \frac{1}{2} \lambda^t H \lambda$.

The dual problem is

(2)

maximize $1^t \lambda - \frac{1}{2} \lambda^t H \lambda$

subject to $\lambda \geq 0$ ← lower bound property

$\lambda^t y = 0$ ← from (4.1)

DUAL PROBLEM

(A standard quadratic optimization problem)

Remark: The matrix H, and therefore the dual problem, depends on the input variable only via the inner products $\langle x_i, x_j \rangle$. This observation will play a major role later when we generalize the approach by making use of kernels; see SL = REPRODUCING KERNEL HILBERT SPACES.

- Analysis of the solution.

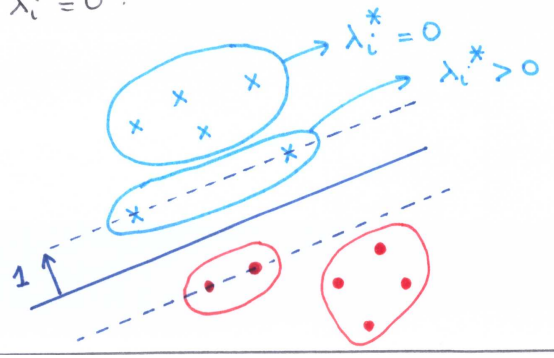
The complementary slackness condition implies that

If $\lambda_i^* > 0$, then $y_i(\beta_0^* + (\beta^*)^t x_i) = 1$

(β_0^*: unspecified for now; $\beta^* = \sum_{i=1}^{n} \lambda_i^* x_i y_i$ (4.2))

$\Rightarrow$ Points (x_i, y_i) with $\lambda_i^* > 0$ lie on the margin.

Conversely, if $y_i(\beta_0^* + (\beta^*)^t x_i) > 1$, then necessarily $\lambda_i^* = 0$.



In addition, we see that only the points (x_i, y_i) with $\lambda_i^* > 0$ determine the nature of the solution $\beta^* = \sum \lambda_i^* x_i y_i$.

These are the SUPPORT VECTORS (SV)

⇒ Once the SVM model is trained, a major proportion of the training sample can be discarded.

× Estimating the threshold.

For any SV, we have

$$y_i(\beta_o^* + (\beta^*)^t x_i) = 1$$
$$\beta_o^* + (\beta^*)^t x_i = y_i$$
$$\beta_o^* = y_i - (\beta^*)^t x_i$$

Pick one to compute β_o^*, or, alternatively, you may average over all SVs, for numerical stability (the optimal λ^* is usually computed up to some degree of accuracy, and so the slackness condition $y_i(\beta_o^* + (\beta^*)^t x_i) = 1$ is only true approximately)

$$\beta_o^* = \frac{1}{|SV|} \sum_{(x_i, y_i) \in SV} \left\{ y_i - \sum_{j \in SV} \lambda_j^* y_j \langle x_i, x_j \rangle \right\}$$

× Remarks (i) Recall that the margin (distance of the SV to the separating hyperplane, on the original scale) is given by $\ell = \frac{1}{\|\beta^*\|}$ (page 12), where

$$\|\beta^*\|^2 = \sum_{i,j \in SV} \lambda_i^* \lambda_j^* y_i y_j \langle x_i, x_j \rangle$$
$$= \sum_{i \in SV} \lambda_i^* \underbrace{\left\{ \sum_{j \in SV} \lambda_j^* y_i y_j \langle x_i, x_j \rangle \right\}}_{= 1 - y_i \beta_o^*}$$

since $y_i(\beta_o^* + (\beta^*)^t x_i) = 1 \quad (i \in SV)$

$$y_i \beta_o^* + \underbrace{\sum_{j \in SV} \lambda_j^* y_i y_j \langle x_i, x_j \rangle} = 1$$

$$\Rightarrow \|\beta^*\|^2 = \sum_{i \in SV} \left\{ \lambda_i^* - \beta_o^* y_i \lambda_i^* \right\} = \sum_{i \in SV} \lambda_i^*$$

since $\sum \lambda_i^* y_i = 0$

Margin is $\boxed{\ell = \frac{1}{\sqrt{\sum_{i \in SV} \lambda_i^*}}}$

(ii) Risk Bound for SVM (linearly separable case)

To derive theoretical guarantees of the SVM algorithm, we introduce the notion of Leave-One-Out error (LOO). Specifically, denoting $f_{\mathcal{L}_n} = \mathcal{A}(\mathcal{L}_n)$ the classifier returned by an algorithm trained on a learning sample $\mathcal{L}_n$, the LOO error on a sample $\mathcal{L}_n$ of size n is

$$\hat{R}_{LOO}(\mathcal{A}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(f_{\mathcal{L}_n \setminus \{(X_i, Y_i)\}}(X_i) \neq Y_i \right)$$

For each i, $\mathcal{A}$ is trained on the learning sample of size (n-1), obtained by removing the i-th observation. Its performance is then evaluated on (X_i, Y_i).

The LOO error satisfies:

$$\mathbb{E}_{\mathcal{L}_n}\left\{ \hat{R}_{LOO}(\mathcal{A}) \right\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{L}_n}\left\{ \mathbb{1}\left(f_{\mathcal{L}_n^{-i}}(X_i) \neq Y_i \right) \right\}$$

$\mathbb{E}\{\dots\}$ under the distribution of $\mathcal{L}_n$.

defining $\mathcal{L}_n^{-i} = \mathcal{L}_n \setminus \{(X_i, Y_i)\}$

$$= \mathbb{E}_{\mathcal{L}_n}\left\{ \mathbb{1}\left(f_{\mathcal{L}_n^{-1}}(X_1) \neq Y_1 \right) \right\}$$
$$= \mathbb{E}\left\{ \mathbb{E}[\dots \mid (X_2, Y_2), \dots, (X_n, Y_n)] \right\}$$
$$= \mathbb{E}_{\mathcal{L}_{n-1}}\left\{ R\left(f_{\mathcal{L}_{n-1}} \right) \right\} \quad \text{---------- (*)}$$

R: 0-1 loss

Identity (*) can be used to prove the following result:

Theorem Let $f_{\mathcal{L}_n}$ = classifier returned by the SVM algorithm on $\mathcal{L}_n$.

$|SV(\mathcal{L}_n)|$ = # of support vectors in $f_{\mathcal{L}_n}$.

Then

$$\mathbb{E}_{\mathcal{L}_n}\{R(f_{\mathcal{L}_n})\} \leq \mathbb{E}_{\mathcal{L}_{n+1}}\left\{\frac{|SV(\mathcal{L}_{n+1})|}{n+1}\right\}$$

The average risk of $f_{\mathcal{L}_n}$ is upper bounded by the average number of SV returned by the SVM algorithm trained on a sample of size $n+1$.

→ Expect that in many cases, there are only a few SVs ⇒ SVM performs well in most cases (⚠ we are still assuming linearly separable data).

proof Note that we have the equivalence

$f_{\mathcal{L}_{n+1}\setminus\{(X_i,Y_i)\}}$ classifies (X_i, Y_i) correctly $\Leftrightarrow$ (X_i, Y_i) is not a SV.

(Indeed, if (X_i, Y_i) is not a SV, then remoing it does not change the expression of the decision boundary, and $f_{\mathcal{L}_n\setminus\{(X_i,Y_i)\}}$ correctly classifies (X_i, Y_i). Conversly, if $f_{\mathcal{L}_n\setminus\{(X_i,Y_i)\}}$ misclassifies (X_i, Y_i), then (X_i, Y_i) must be a SV.

Thus $\hat{R}_{LOO}(A) \leq \frac{|SV(\mathcal{L}_{n+1})|}{n+1}$ + take expectation + (*) ∎

[Ref] Mohri & Al. Foundations of Machine Learning.

II.2. The linearly non-separable case.

We generalise the SVM algorithm to allow the classification of linearly non-separable points, by allowing some points to be on the wrong side of the margin.

Specifically, for a linearly non-separable training set, there exists $(x_i, y_i) \in \mathcal{L}_n$ such that $y_i(\beta_0 + \beta^t x) \not\geq 1$

↑ Distance of (x_i, y_i) to the hyperplane $\{x \mid \beta_0 + \beta^t x = 0\}$ is not ≥ 1.

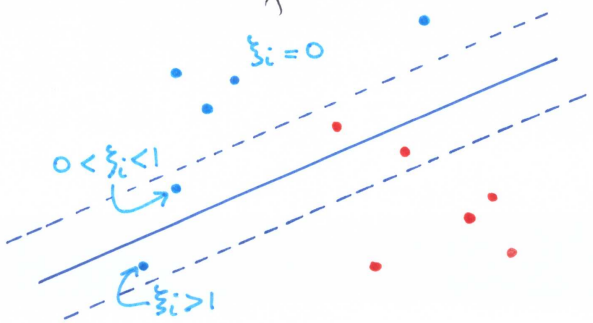
⇒ We relax this constraint by introducing SLACK VARIABLES $\xi_i \geq 0$, $1 \leq i \leq n$, providing the extra room needed

↑ one for each (x_i, y_i)

to perform classification: ξ_i measures the distance by which observation x_i violates the condition $y_i(\beta_0 + \beta^t x_i) \geq 1$.

Precisely, ξ_i is chosen as the smallest positive number such that $y_i(\beta_0 + \beta^t x_i) + \xi_i \geq 1$. We have

- $\xi_i = 0$: points on the correct side of the margin
- $0 < \xi_i \leq 1$: margin violate; but x_i lies on the correct side
- $\xi_i = 1$: x_i lies on the decision boundary
- $\xi_i > 1$: observation x_i is misclassified



The goal is to maximize the margin, while soft penalizing points that lie on the wrong side of the boundary: $\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i$

↑ Trade-off between these two terms:

- Small C: the constraint is easily ignored ⇒ margin is large
- Large C: narrow margin

margin is defined as $\ell := \frac{1}{\|\beta\|}$ →

Note that a misclassified point has $\xi_i > 1$. The sum $\sum_{i=1}^{n}\xi_i$ represent an upper bound on the number of misclassified points. In Section II.4, we introduce an alternative approach to this optimization problem, a more natural interpretation.

(3)

minimize $\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i$

subject to $y_i(\beta_0 + \beta^t x_i) \geq 1 - \xi_i$ $\quad 1 \leq i \leq n$

$\xi_i \geq 0$

PRIMAL PROBLEM.

↳ Lagrangian is

$$\mathcal{L}(\beta_0, \beta, \xi, \lambda, \nu) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\nu_i\xi_i - \sum_{i=1}^{n}\lambda_i\{y_i(\beta_0 + \beta^t x_i) - 1 + \xi_i\}$$

($\lambda, \nu \in \mathbb{R}^n$)

↳ Lagrange dual function is $g(\lambda, \nu) = \inf_{\beta_0, \beta, \xi} \mathcal{L}(\beta_0, \beta, \xi, \lambda, \nu)$.

Strong duality holds for this optimization problem.

⇒ The primal optima $\beta_0^*, \beta^*, \xi^*$ & the dual optima satisfy the KKT conditions.

• KKT conditions (non-linearly separable case)

① Primal Constraints $\quad 1 - y_i(\beta_0 + \beta^t x_i) - \xi_i \leq 0$

$-\xi_i \leq 0$

② Dual Constraints $\quad \lambda_i \geq 0$

$\nu_i \geq 0$

③ Complementary Slackness $\quad \lambda_i(y_i(\beta_0 + \beta^t x_i) - 1 + \xi_i) = 0$

$\nu_i \xi_i = 0$

④ Gradient w.r.t. β_0, β, ξ, of the Lagrangian vanishes

$$\frac{\partial \mathcal{L}(\beta_0, \beta, \xi)}{\partial \beta_0} = -\sum \lambda_i y_i = 0 \qquad (4.1)$$

$$\frac{\partial \mathcal{L}(\beta_0, \beta, \xi)}{\partial \beta} = \beta - \sum \lambda_i y_i x_i = 0 \qquad (4.2)$$

$$\frac{\partial \mathcal{L}(\beta_0, \beta, \xi)}{\partial \xi} = C - \nu - \lambda = 0 \qquad (4.3)$$

↑ We proceed as for the linearly separable case: eliminate β_0, β & ξ from the expression of the Lagrangian, using (4.1) – (4.3).

$$\underbrace{\frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle}_{\frac{1}{2}\|\beta\|^2} + C\sum_i \xi_i - \sum \lambda_i y_i \beta^t x_i - \sum \lambda_i y_i \beta_0 + \sum \lambda_i - \sum \lambda_i \xi_i - \sum (C - \lambda_i)\xi_i$$

(4.2): $\beta = \sum \lambda_i y_i x_i$

(4.3): $\nu_i = C - \lambda_i$

↳ The Lagrange dual function is

$$g(\lambda) = -\frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^{n}\lambda_i$$

$$= -\frac{1}{2}\lambda^t H \lambda + 1^t\lambda$$

• $1 = \begin{pmatrix}1\\1\\1\end{pmatrix}$ $\lambda = \begin{pmatrix}\lambda_1\\ \vdots \\ \lambda_n\end{pmatrix}$

• $H = \left(y_i y_j \langle x_i, x_j\rangle\right)$ ← (as on page 13)

Constraints on λ_i are : (2) $\lambda_i \geq 0$
(4.3) $\nu_i = C - \lambda_i \geq 0$ } $0 \leq \lambda_i \leq C$
(4.1) $1^t \lambda = 0$

$\Rightarrow$

maximize $1^t\lambda - \frac{1}{2}\lambda^t H \lambda$ (4)
subject to $\lambda^t y = 0$
$0 \leq \lambda \leq C$ ← "box constraint"

DUAL PROBLEM

We discuss in section II.3 the SMO algorithm to solve (4).
Once the dual optimum λ^* (& ν^*) are computed, we get
$\beta^* = \sum_{i=1}^{n} \lambda_i^* y_i x_i$ (4.2).

• Analysis of the solution.

x If $\lambda_i^* > 0$, then $y_i(\beta_0^* + (\beta^*)^t x_i) = 1 - \xi_i$ (3)
↳ If in addition $\lambda_i^* < C$, then $\nu_i > 0$ & $\xi_i = 0$ (4.2) (3)
Points lie on the margin

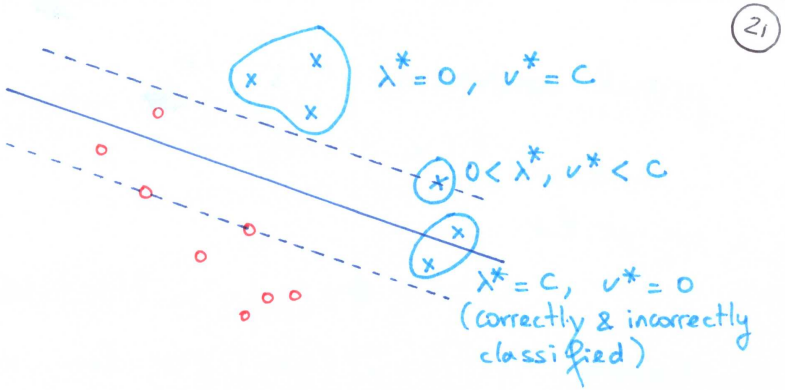
[KKT.1] $0 < \lambda_i^* < C \Leftrightarrow 0 < \nu_i^* < C \Leftrightarrow y_i(\beta_0^* + (\beta^*)^t x_i) = 1$

x If $\lambda_i^* = C$, then $y_i(\beta_0^* + (\beta^*)^t x_i) = 1 - \xi_i \leq 1$ (3)
& $\nu_i^* = 0$ (4.3)

[KKT.2] $\lambda_i^* = C \Leftrightarrow \nu_i^* = 0 \Leftrightarrow y_i(\beta_0^* + (\beta^*)^t x_i) \leq 1$

x If $\nu_i^* = C$, then $\lambda_i^* = 0$ & $\xi_i^* = 0$
and (1) $\Rightarrow y_i(\beta_0^* + (\beta^*)^t x_i) \geq 1$ (3) : points away from the boundary.

[KKT.3] $\lambda_i^* = 0 \Leftrightarrow \nu_i^* = C \Leftrightarrow y_i(\beta_0^* + (\beta^*)^t x_i) \geq 1$



Observations with $0 < \lambda_i^* < C$ lie on the margin. These can be used to determine the intercept since $y_i(\beta_0^* + (\beta^*)^t x_i) = 1$
$0 < \nu_i^* < C$

$\beta_0^* = y_i - (\beta^*)^t x_i$.

For numerical stability, we average over all points lying on the 'edge' (denote this set of points SV_e)

$$\beta_0^* = \frac{1}{|SV_e|} \sum_{i \in SV_e} \left\{ y_i - \sum_{j \in SV_e} \lambda_j^* y_j \langle x_i, x_j \rangle \right\}$$

x Remarks. (i) ERM view of SVM.

An analysis of the SVM solution shows that points that lie on the correct side of the margin (those for which $y_i(\beta_0^* + (\beta^*)^t x_i) \geq 1$) have $\xi_i = 0$ ([KKT.1] & [KKT.3]). The remaining points are such that $\xi_i = 1 - y_i(\beta_0^* + (\beta^*)^t x_i)$ ([KKT.2]). The optimization problem (3) is thus equivalent to

$$\underset{\beta_0, \beta}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \left(1 - y_i\{\beta_0 + \beta^t x_i\}\right)_+ + \lambda \|\beta\|^2$$

↑ The hinge loss

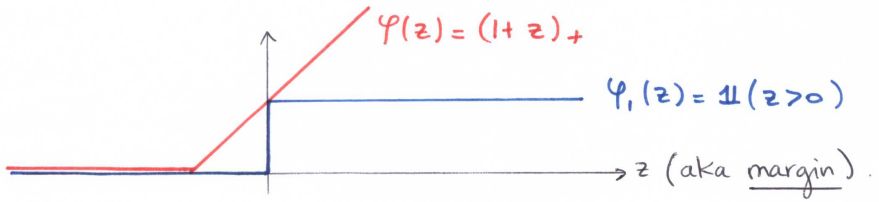
$\Leftrightarrow$ minimize$_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|\beta\|^2 \right\}$,

$\mathcal{F} = \{ x \mapsto \beta_0 + \beta^t x \}$

where $\ell(y_i, f(x_i)) = (1 - y_i f(x_i))_+ = \text{hinge loss}$.

$= \varphi(-y_i f(x_i))$, $\varphi(z) = \max(0, 1+z)$

convex surrogate to the 0-1 loss, see SL = CONVEX RELAXATION



(ii) Margin Theory.

SVM can be motivated as well from a more theoretical viewpoint, using elements of "margin theory". Specifically, one can show that for $\mathcal{F} = \{ x \mapsto \beta^t x \; ; \; \|\beta\| \leq \Lambda \}$

$\forall f \in \mathcal{F}$, $R(f) \leq \underbrace{\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + 2\sqrt{\frac{r^2 \Lambda^2}{n}}}_{(*)} + \sqrt{\frac{\log 1/\delta}{2n}}$

$R(f) = \mathbb{P}(Yf(X) \leq 0)$

assuming $\|x\| \leq r$

with probability $\geq 1 - \delta$

(See expression (4.47) in Mohri et al).

The SVM optimization problem is precisely the minimization of this upper bound, since minimizing Λ (recall $\|\beta\| \leq \Lambda$) is equivalent to minimizing $\|\beta\|$ or $\|\beta\|^2$.

(iii) For multi-class classification problems, the R package 'e1071' is implementing a 'one-vs-one' approach → $K(K-1)/2$ binary classifiers are trained; and then a voting scheme is used.

Alternatively, we may directly generalize the expression of the hinge loss, and adapt it to the multi-class problem (the one vs one & one vs all approaches ignore the nature of the multiclass problem, by reducing it to a binary classification task) → we discuss this in Section I.5.

(iv) An online-version of SVM is presented in SL = GDA, using a stochastic gradient algorithm → Algorithm "Pegasos" (Shalev-Schwartz, Singer, Srebro, Cotter (2011)).

I.3. Sequential Minimal Optimization (SMO).

The SMO algorithm is an algorithm that can solve the dual problem (4) (page 20):

maximize $1^t \lambda - \frac{1}{2} \lambda^t H \lambda$

subject to $\lambda^t y = 0$

$0 \leq \lambda \leq C$

SMO solves the maximization problem recursively, involving two Lagrange multipliers λ_i and λ_j at a time.

→ Cannot update only one multiplier at a time, since we have the linear constraint $\lambda^t y = 0 \Leftrightarrow \lambda_i y_i = -\sum_{k \neq i} \lambda_k y_k$

$y_i^2 = 1$ ↪ $\lambda_i = -\underbrace{\sum_{k \neq i} \lambda_k y_k y_i}$

fixed when holding λ_k fixed $(k \neq i)$

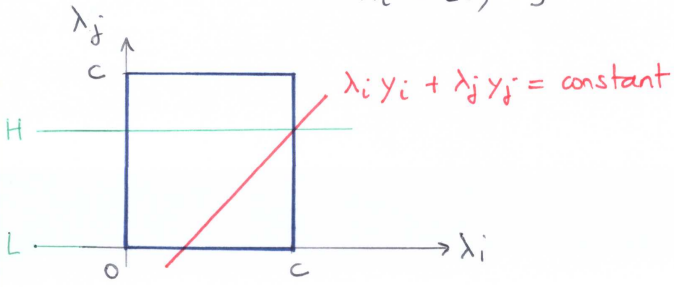
→ Once two Lagrange multipliers are selected, their update is relatively straightforward, since making use of the linear constraint $\lambda_i y_i + \lambda_j y_j = -\sum_{k \neq i,j} \lambda_k y_k$ and keeping the remaining λ_k fixed ($k \neq i, j$), we can express λ_j as a function of λ_i, and solve for λ_j, as the minimizer of a quadratic function.

↳ The choice of λ_i & λ_j is made using heuristics. Platt suggests picking those that will yield the largest step towards the global minimum. We may consider a simplified procedure, and iterate over all values of λ_i until we find one that violates the KKT conditions (KKT. 1, KKT. 2 & KKT. 3 page 20). Once such a λ_i is selected, select λ_j randomly amongst the remaining Lagrange multipliers. Repeat until all λs satisfy the KKT conditions, within some tolerance.

Suppose we picked λ_i and λ_j, and keep λ_k fixed ($k \neq i, j$).

↳ Then $\lambda_i y_i + \lambda_j y_j = -\sum_{k \neq i,j} \lambda_k y_k = \text{constant}$.

↖ There are constraints on the values of λ_i and λ_j, to ensure that both of them lie in the interval $[0, C]$: We derive bounds L and H on λ_j; $L \leq \lambda_j \leq H$, to ensure that $\lambda_i \in [0, C]$



• Consider two cases:

• If $y_i = y_j$, then $\lambda_i + \lambda_j = -\sum_{k \neq i,j} \lambda_k y_k y_i = \alpha$ (a constant)

$$0 \leq \lambda_i \leq C$$
$$0 \leq \alpha - \lambda_j \leq C \quad \hookleftarrow \lambda_i + \lambda_j = \alpha$$
$$-C \leq \lambda_j - \alpha \leq 0$$
$$\alpha - C \leq \lambda_j \leq \alpha$$
$$\lambda_i + \lambda_j - C \leq \lambda_j \leq \lambda_i + \lambda_j \quad \& \quad 0 \leq \lambda_j \leq C$$

$$\max(0, \lambda_i + \lambda_j - C) \leq \lambda_j \leq \min(C, \lambda_i + \lambda_j)$$

• If $y_i \neq y_j$, then $\lambda_i - \lambda_j = -\sum_{k \neq i,j} \lambda_k y_k y_i = \alpha$

$$0 \leq \lambda_i \leq C$$
$$0 \leq \alpha + \lambda_j \leq C$$
$$-\alpha \leq \lambda_j \leq C - \alpha$$
$$\lambda_j - \lambda_i \leq \lambda_j \leq C + \lambda_j - \lambda_i \quad \& \quad 0 \leq \lambda_j \leq C$$

$$\max(0, \lambda_j - \lambda_i) \leq \lambda_j \leq \min(C, C + \lambda_j - \lambda_i)$$

• Conclusion,

If $y_i = y_j$, $L = \max(0, \lambda_i + \lambda_j - C)$, $H = \min(C, \lambda_i + \lambda_j)$

If $y_i \neq y_j$, $L = \max(0, \lambda_j - \lambda_i)$, $H = \min(C, C + \lambda_j - \lambda_i)$

↳ Then express λ_i^{old} as a function of λ_j^{old}:

$$\lambda_i^{old} = -\sum_{k \neq i} \lambda_k^{old} y_k y_i = -\lambda_j^{old} y_j y_i - \text{constant}$$

↳ Maximize a quadratic function with respect to λ_j only. Since the solution $\tilde{\lambda}_j$ is not guaranteed to lie

in the interval $[L, H]$, we then "clip" it to force it to be in it.

$$\lambda_j^{new} = \begin{cases} H & \text{if } \tilde{\lambda}_j > H \\ \tilde{\lambda}_j & \text{if } L \le \tilde{\lambda}_j \le H \\ L & \text{if } \tilde{\lambda}_j < L \end{cases}$$

We then compute $\lambda_i^{new} = \lambda_i^{old} + y_i y_j (\lambda_j^{old} - \lambda_j^{new})$

Since $\lambda_i^{old} = -\lambda_j^{old} y_j y_i$ - constant

add $\lambda_j^{old} y_j y_i$ & substract $\lambda_j^{new} y_j y_i$ to get λ_i^{new}.

- Once λ_i^{new} and λ_j^{new} are computed, we update the value of the threshold β_0 such that the KKT conditions are satisfied for these 2 Lagrange multipliers.

→ If $0 < \lambda_i^{new} < C$, then [KKT.1] $\Rightarrow \beta_0 = y_i - \beta_{new}^t x_i$, where

$$\beta_{new} = \sum_{k=1}^{n} \lambda_k^{new} y_k x_k$$
$$= \sum_{k=1}^{n} \lambda_k^{old} y_k x_k + y_i x_i (\lambda_i^{new} - \lambda_i^{old}) + y_j x_j (\lambda_j^{new} - \lambda_j^{old})$$

$$\Rightarrow \beta_0^i \leftarrow y_i - \beta_{old}^t x_i - y_i (\lambda_i^{new} - \lambda_i^{old}) \langle x_i, x_i \rangle - y_j (\lambda_j^{new} - \lambda_j^{old}) \langle x_i, x_j \rangle$$

→ If $0 < \lambda_j^{new} < C$ as well, then

$$\beta_0^j \leftarrow y_j - \beta_{old}^t x_j - y_i (\lambda_i^{new} - \lambda_i^{old}) \langle x_i, x_j \rangle - y_j (\lambda_j^{new} - \lambda_j^{old}) \langle x_j, x_j \rangle$$, and $\beta_0^j = \beta_0^i = \beta_0^{new}$

→ If $\lambda_i^{new} = 0$ or C and $\lambda_j^{new} = 0$ or C, then any value between β_0^i and β_0^j satisfy the KKT conditions, and we take $\beta_0^{new} \leftarrow (\beta_0^i + \beta_0^j)/2$.

Summary: $\beta_0^{new} = \begin{cases} \beta_0^i & \text{if } 0 < \lambda_i^{new} < C \\ \beta_0^j & \text{if } 0 < \lambda_j^{new} < C \end{cases} = \frac{\beta_0^i + \beta_0^j}{2}$ in any other case

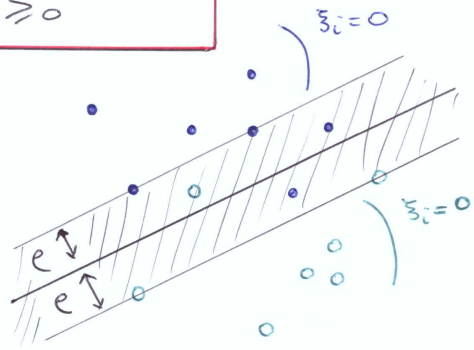
II.4. ν-SV classification.

The original optimization problem (page 18) in the linearly non-separable case contains a tuning parameter C, chosen a-priori, or using a validation technique. We substitute C by a new parameter $\nu \ge 0$, which, as we shall see, has a more natural interpretation. We consider the following optimization problem (Schölkopf et al (2000)):

$$\begin{aligned} \underset{\beta_0, \beta, \varrho, \xi}{\text{minimize}} \quad & \frac{1}{2}\|\beta\|^2 - \nu\varrho + \sum \xi_i \\ \text{subject to} \quad & y_i(\beta_0 + \beta^t x_i) \ge \varrho - \xi_i \\ & \xi_i \ge 0 \\ & \varrho \ge 0 \end{aligned}$$

← constant C disappears

ϱ plays the role of the margin: points with $\xi_i = 0$ are separated by the margin 2ϱ.



- Lagrangian is

$$L(\beta_0, \beta, \xi, \varrho, \lambda, \mu, \delta) = \frac{1}{2}\|\beta\|^2 - \nu\varrho + \sum \xi_i - \sum_{i=1}^{n} \lambda_i [y_i f(x_i) - \varrho + \xi_i] - \sum_{i=1}^{n} \mu_i \xi_i - \delta\varrho$$

where $f(x_i) = \beta_0 + x_i^t \beta$.

KKT conditions.

(1) Primal Constraints $\quad y_i(\beta_0 + \beta^t x_i) \geq e - \xi_i$
$\xi_i \geq 0$
$e \geq 0$

(2) Dual Constraints $\quad \lambda, \mu, \delta \geq 0$

(3) Complementary Slackness $\quad \lambda_i [y_i f(x_i) - e + \xi_i] = 0$
$\mu_i \xi_i = 0$
$\delta e = 0$

(4) Gradient of Lagrangian vanishes:

- $\frac{\partial L}{\partial \beta_0} = -\sum \lambda_i y_i = 0$ ——— (4.1)
- $\frac{\partial L}{\partial \beta} = \beta - \sum_{i=1}^{n} \lambda_i y_i x_i = 0$ ——— (4.2)
- $\frac{\partial L}{\partial \xi_i} = 1 - \lambda_i - \mu_i = 0$ ——— (4.3)
- $\frac{\partial L}{\partial e} = -\upsilon + \sum \lambda_i - \delta = 0$ ——— (4.4)

- **Dual problem.** We get from (4.2) that $\beta(\lambda) = \sum \lambda_i y_i x_i$, and from (4.4) that $\delta = \sum \lambda_i - \upsilon$. Also, (4.3) ensures that $\mu = 1 - \lambda$. We obtain

$L(\beta_0, \beta(\lambda), \xi, e, \lambda, (1-\lambda), (\sum \lambda_i - \upsilon)) =$ ($(1-\lambda) = \mu$, $(\sum \lambda_i - \upsilon) = \delta$)

$\frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^t x_j - \upsilon e + \sum \xi_i$ $\quad = -\sum \lambda_i y_i f(x_i)$

$\frac{1}{2}\|\beta(\lambda)\|^2$ $\quad - \sum_{i=1}^{n} \lambda_i y_i \beta_0 - \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^t x_j$

$+ e \sum \lambda_i - \sum \lambda_i \xi_i - \sum (1-\lambda_i) \xi_i$

$- (\sum \lambda_i - \upsilon) e$

We obtain $g(\lambda) = -\frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^t y_j$

$= -\frac{1}{2} \lambda^t H \lambda$, using matrix notations.

The dual problem maximizes g under the constraints:

- $\lambda \gtrless 0$
- $\mu \gtrless 0 \Rightarrow \lambda \lesssim 1$
- $\delta \geq 0 \Rightarrow \sum \lambda_i \geq \upsilon$
- $\sum \lambda_i y_i = 0$

Again, the kernel trick can be applied here

DUAL PROBLEM.

maximize	$-\frac{1}{2} \lambda^t H \lambda$
subject to	$0 \lesssim \lambda \lesssim 1$
	$\sum \lambda_i \geq \upsilon$
	$1^t \lambda = 0$

— Denote by λ^* the solution of the dual —
Once λ^* is computed, $\mu^* = 1 - \lambda^*$ and $\delta^* = \sum \lambda_i^* - \upsilon$ follow immediately.
Moreover, $\beta^* = \sum \lambda_i^* y_i x_i$ ⇒ only observations with $\lambda_i^* > 0$ contribute to the solution (aka the support vectors).

- **Expressions for β_0^* and e^***

Let $\mathcal{S}^+ = \{ \text{obs} \mid 0 < \lambda_i^* < 1 \text{ and } y_i = +1 \}$
$\mathcal{S}^- = \{ \text{obs} \mid 0 < \lambda_i^* < 1 \text{ and } y_i = -1 \}$

and consider $S^+ \subset \mathcal{S}^+$
$S^- \subset \mathcal{S}^-$, such that S^+ and S^- are of equal size $|S^+| = |S^-| = S$

Then

$$\left.\begin{aligned} \forall i \in S^+, \quad & \beta_0^* + x_i^t \beta^* = e^* \\ \forall i \in S^-, \quad & \beta_0^* + x_i^t \beta^* = -e^* \end{aligned}\right\} (*)$$

Indeed, for points in S^+/S^-, $0 < \lambda_i^* < 1$

$\Rightarrow 0 < \mu_i^* < 1$

$\Rightarrow \xi_i = 0$ from complementary slackness

Thus $\underbrace{\lambda_i}_{>0} [\, y_i f(x_i) - \epsilon + \underbrace{\xi_i}_{=0} \,] = 0 \Rightarrow y_i f(x_i) = \epsilon$.

Summing (*) over all observations in S^+/S^- yields

$$s\,\beta_0^* + \sum_{i \in S^+} \sum_j \lambda_j^* y_j x_i^t x_j = s\,\epsilon^*$$

$$s\,\beta_0^* + \sum_{i \in S^-} \sum_j \lambda_j^* y_j x_i^t x_j = -s\,\epsilon^*$$

---

sum $\Rightarrow$

$$\beta_0^* = \frac{1}{2s} \sum_{i \in S^+ \cup S^-} \sum_j \lambda_j^* y_j x_i^t x_j$$

difference $\Rightarrow$

$$\epsilon^* = \frac{1}{2s} \left\{ \sum_{i \in S^+} \sum_j \lambda_j^* y_j x_i^t x_j - \sum_{i \in S^-} \sum_j \lambda_j^* y_j x_i^t x_j \right\}$$

Again, sums over S^+/S^- are considered for practical purposes

- Consequences: Having computed β_0^* and β^*, we classify a new point according to

$$\hat{f}(x) = \text{sign}\left(\beta_0^* + x^t \beta^* \right).$$

Suppose now that the original optimization problem includes the constant C:

$$\begin{aligned} \text{maximize} \quad & \frac{1}{2}\|\beta\|^2 + C\left(-\nu\epsilon + \sum \xi_i\right) \\ \text{s.t.} \quad & y_i(\beta_0 + \beta^t x_i) \geq \epsilon - \xi_i \\ & \xi_i \geq 0 \\ & \epsilon \geq 0 \end{aligned}$$

Proceed as before. It can easily be seen that the expression of the lagrange dual function remains unchanged, and that only the dual constraints are modified to:

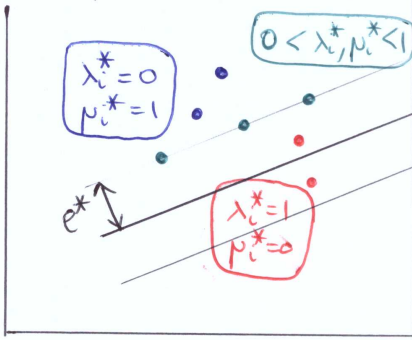
$$\begin{aligned} & 0 \leq \lambda \leq C \\ & \textstyle\sum \lambda_i \geq C\nu \\ & 1^t \lambda = 0 \end{aligned}$$

$\Rightarrow$ The solution of the dual problem is scaled by C, and the corresponding decision function $\hat{f}(x)$ does not change (check the expression of β_0^* and β^*).

$\Rightarrow$ We can take $C = 1$ without loss of generality.

- Analysis of the solution.

Suppose that you run ν-SV on some data with the result that $\epsilon^* > 0$ (note that the primal constraint ensure $\epsilon^* \geq 0$, always). Complementary slackness then implies that $\delta^* = 0$, which gives $\sum \lambda_i^* = \nu$.



(i) Points outside the margin are such that $y_i f(x_i) > \epsilon^*$. Complementary Slackness $\Rightarrow$

$\lambda_i^* [\, \underbrace{y_i f(x_i) - \epsilon^* + \xi_i}_{>0} \,] = 0$

$\Rightarrow \lambda_i^* = 0$ & $\mu_i^* = 1$

(ii) Suppose now that $\lambda_i^* > 0$. Then $y_i f(x_i) = \epsilon^* - \xi_i$. If in addition $0 < \lambda_i^* < 1$, then $0 < \mu_i^* < 1$ and thus from complementary slackness, we get $\xi_i = 0 \Rightarrow$ points on the margin

(iii) Points within the margin have $\xi_i > 0$, so that $\mu_i^* = 0$, and $\lambda_i^* = 1$.

- Further observations

Suppose here as well that after running the ν-SV on some data, you obtain $\varrho^* > 0$. This implies that $\sum \lambda_i^* = \nu$.

Since the support vectors (ie. the observations for which $\lambda_i^* > 0$) necessarily have $0 < \lambda_i^* \leq 1$, the sum $\sum \lambda_i^*$ represents a lower bound on the number of SV.

$\Rightarrow$ ν is a lower bound on the number of support vectors.

Consider the points that strictly lie in the margin:

$$N := \{ \text{observations} \mid y_i f(x_i) < \varrho \}$$

Then a similar reasoning shows that necessarily this number of points N must be less that $\sum \lambda_i^* = \nu$ (since all points in the margin have $\lambda_i^* = 1$)

$\Rightarrow$ ν is an upper bound on the number of points strictly within the margin.

II.5. Multiclass SVM.

- Recall that the hinge loss in the binary classification problem is

$$\ell(y, f(x)) = \max(0, 1 - y f(x)), \quad y \in \{-1, +1\}.$$

In the context of SVM, $f(x) = \beta_0 + \beta^t x$

$=$ linear function of $x \in \mathbb{R}^d$

We write $f(x) = \langle \beta, x \rangle$, incorporating the constant into the vector β, and augmenting x, so that $x \in \mathbb{R}^{d+1}$

- Instead of implementing a one-vs-one or one-vs-all strategy to use SVM in a K-class classification task, we propose here to consider the multiclass nature of the problem by directly generalizing the definition of the hinge loss.

- A general strategy, already encountered when discussing LDA, is to consider a family of discriminant functions $\delta_1(x), \dots, \delta_K(x)$, and to classify x according to

$$\hat{y} \in \underset{1 \leq k \leq K}{\operatorname{argmax}} \ \delta_k(x) \qquad \hat{y} \in \{1, 2, \dots, K\}$$

We consider linear discriminants:

$$\delta_k(x) = \beta_{k1} x_1 + \cdots + \beta_{kd} x_d = \langle b_k, x \rangle,$$

and put

$$B = \begin{pmatrix} \text{—} & b_1 & \text{—} \\ \text{—} & b_2 & \text{—} \\ & \vdots & \\ \text{—} & b_K & \text{—} \end{pmatrix} \quad (K \times d)$$

The goal is to define a function $\ell(B, (x, y))$ which generalizes the hinge loss $\max(0, 1 - \langle \beta, x \rangle)$ in the binary case. In addition, the multiclass hinge loss should be a convex surrogate for the 0-1 loss

$$\mathbb{1}(\hat{y} \neq y) \qquad \hat{y}, y \in \{1, 2, \dots, K\}$$

Note that $\forall y$

$$\mathbb{1}(\hat{y} \neq y) \leq \max_j \{ \mathbb{1}(j \neq y) + \langle b_j, x \rangle - \langle b_y, x \rangle \}$$

(*)

Indeed, if $y = \hat{y}$, the LHS $= 0$, while the RHS is always non-negative, as $j = \hat{y} = y$ ensures that RHS ≥ 0. Moreover, if $y \neq \hat{y}$, then LHS $= 1$, and

$$\text{RHS} \geq \mathbb{1}(y \neq \hat{y}) + \underbrace{\langle b_{\hat{y}}, x \rangle - \langle b_y, x \rangle}_{\geq 0} \geq 1.$$

↑ take $j = \hat{y}$ — by definition of $\hat{y}$

- We define

$$\ell(B, (x,y)) := \max_j \left\{ \mathbb{1}(j \neq y) + \langle b_j, x \rangle - \langle b_y, x \rangle \right\}$$

MULTICLASS HINGE LOSS.

- Let's try to understand why it is meaningful to call this function the Multiclass Hinge loss.

First of all, in the case of binary classification, taking

$$B = \frac{1}{2} \begin{pmatrix} \text{—} & b & \text{—} \\ \text{—} & -b & \text{—} \end{pmatrix} \quad (2 \times d)$$

$\leftarrow b_1 = \frac{b}{2}$

$\nwarrow b_{-1} = -\frac{b}{2}$

(consider class $+1$, and class -1)

Then

$$\ell(B, (x,y)) = \max \left\{ \mathbb{1}(y \neq 1) + \langle b_1, x \rangle - \langle b_y, x \rangle, \; \mathbb{1}(y \neq -1) + \langle b_{-1}, x \rangle - \langle b_y, x \rangle \right\}$$

($j = +1$ and $j = -1$)

$$= \max \left\{ \mathbb{1}(y = -1) + \frac{1}{2}\langle b, x \rangle - \langle b_y, x \rangle, \; \mathbb{1}(y = +1) - \frac{1}{2}\langle b, x \rangle - \langle b_y, x \rangle \right\}$$

If $y = +1$, then

$$\ell(B, (x,y)) = \max \{ 0, \; 1 - \langle b, x \rangle \} = \max \{ 0, \; 1 - y\langle b, x \rangle \}$$

If $y = -1$, then

$$\ell(B, (x,y)) = \max \{ 1 + \langle b, x \rangle, \; 0 \} = \max \{ 1 - y\langle b, x \rangle, \; 0 \}.$$

In both cases, $\ell(B, (x,y)) = \max \{ 0, 1 - y\langle b, x \rangle \}$

$$= \text{hinge loss}$$

↳ The multiclass hinge loss reduces to the usual binary hinge loss in the binary classification setting. Recall that the binary hinge loss incurs no loss for observations that are at (absolute) distance at least $+1$ from the decision boundary:

$$\max(0, 1 - y\langle b, x \rangle) = 0 \quad \text{if} \quad y\langle b, x \rangle \leq 1$$

$$\Updownarrow$$

$$1 - y\langle \tfrac{b}{2}, x \rangle \leq y\langle \tfrac{b}{2}, x \rangle$$

If $y = +1$, then no loss if $1 - \langle \frac{b}{2}, x \rangle \leq \langle \frac{b}{2}, x \rangle$

i.e. $1 + \langle b_{-1}, x \rangle \leq \langle b_1, x \rangle$

If $y = -1$, then no loss if $1 + \langle \frac{b}{2}, x \rangle \leq -\langle \frac{b}{2}, x \rangle$

i.e. $1 + \langle b_1, x \rangle \leq \langle b_{-1}, x \rangle$

↳ Putting things together, the binary hinge loss incurs no cost if for $j \neq y$, we have

$$1 + \langle b_j, x \rangle \leq \langle b_y, x \rangle$$

i.e. if the correct class has margin $+1$ with the other class.

The same is true for the multiclass hinge loss : if

$\forall j \neq y, \quad 1 + \langle b_j, x \rangle \leq \langle b_y, x \rangle,$

then

$$\ell(B, (x,y)) = \max_j \{ \mathbb{1}(j \neq y) + \langle b_j, x \rangle - \langle b_y, x \rangle \}$$
$$= \max \left\{ \max_{j \neq y} \underbrace{(1 + \langle b_j, x \rangle - \langle b_y, x \rangle)}_{\leq 0}, 0 \right\}$$
$$\leq 0$$

⇒ No loss occurs if the correct class y has margin $+1$ with all other classes.

↳ All these observations justify the name "multiclass hinge loss" for the function $\ell(B, (x,y))$ thus defined.

⇒ Given $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$, the multi-class SVM optimization problem is

$$\underset{W}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(W, (x_i, y_i)) + \frac{\lambda}{2} \|W\|^2$$

Frobenius norm

$$\|W\|^2 = \sum_{k=1}^{K} \sum_{j=1}^{d} |w_{kj}|^2$$

Minimization can be achieved for example using a stochastic gradient algorithm, see SL= GRADIENT DESCENT ALGORITHMS

[Ref] S. Shalev-Schwartz & S. Ben-David. Understanding Machine Learning.

References

Cortes C., Vapnik V. (1995). Support Vector Networks. Machine Learning, vol 20, 273-297.

Schölkopf B, Smola A.J., Williamson R.C., Bartlett, P.L. (2000). New Support Vector Algorithms. Neural Computation, vol 12, 1207-1245

Platt J.C. (1998) Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Machines. Advances in Kernel Methods - Support Vector Learning

Shalev-Schwartz S., Singer Y., Srebro N., Cotter A. (2011). Pegasos: Primal Estimated Sub-Gradient Solver for SVM. Mathematical Programming, vol 127, No 1, p. 3-30.