Introduction to Machine Learning: Kernels

Part 2: Convex optimization, support vector machines

Arthur Gretton

Gatsby Unit, CSML, UCL

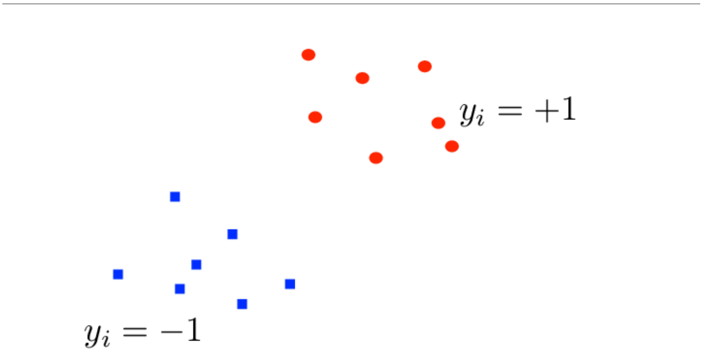
May 23, 2017

Overview

- Review of convex optimization
- Support vector classification, the C-SV machine

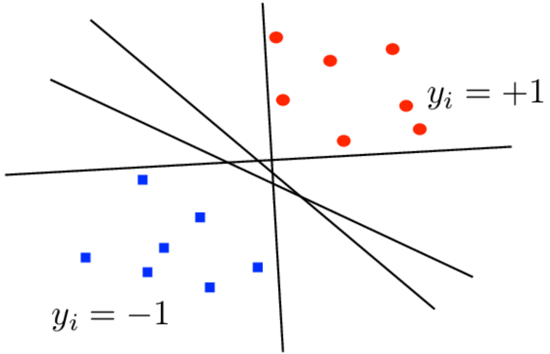
Why we need optimization: SVM idea

Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



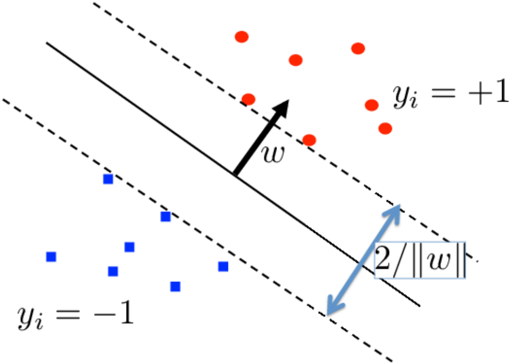
Why we need optimization: SVM idea

Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



Why we need optimization: SVM idea

Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



Smallest distance from each class to the separating hyperplane $w^\top x + b$ is called the **margin**.

Why we need optimization: SVM idea

This problem can be expressed as follows:

$$\max_{w,b}(\mathrm{margin}) = \max_{w,b}\left(\frac{2}{\|w\|}\right) \quad \text{or} \quad \min_{w,b}\|w\|^2 \tag{1}$$

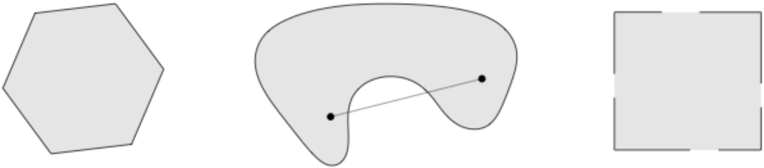
subject to

$$\begin{cases} w^\top x_i + b \geq 1 & i\ :\ y_i = +1, \\ w^\top x_i + b \leq -1 & i\ :\ y_i = -1. \end{cases} \tag{2}$$

This is a convex optimization problem.

Short overview of convex optimization

Convex set



(Figure from Boyd and Vandenberghe)

Leftmost set is convex, remaining two are not.
Every point in the set can be seen from any other point in the set, along a straight line that never leaves the set.

Definition

C is convex if for all $x_1, x_2 \in C$ and any $0 \leq \theta \leq 1$ we have $\theta x_1 + (1-\theta)x_2 \in C$, i.e. every point on the line between x_1 and x_2 lies in C.

Convex function: no local optima

(Figure from Boyd and Vandenberghe)

Definition (Convex function)

A function f is **convex** if its domain $\mathrm{dom} f$ is a convex set and if $\forall x, y \in \mathrm{dom} f$, and any $0 \leq \theta \leq 1$,

$$f\left(\theta x + (1-\theta)y\right) \leq \theta f(x) + (1-\theta) f(y).$$

The function is **strictly convex** if the inequality is strict for $x \neq y$.

Optimization and the Lagrangian

Optimization problem on $x \in \mathbb{R}^n$,

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0 \qquad i = 1, \ldots, m \end{aligned} \tag{3}$$

- p^* the optimal value of (3)

Ideally we would want an unconstrained problem

$$\text{minimize } f_0(x) + \sum_{i=1}^{m} I_{-}\left(f_i(x)\right),$$

where $I_{-}(u) = \begin{cases} 0 & u \leq 0, \\ \infty & u > 0. \end{cases}$

Why is this hard to solve?

Optimization and the Lagrangian

Optimization problem on $x \in \mathbb{R}^n$,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0 \qquad\qquad i = 1, \ldots, m \end{array} \tag{3}$$

- p^* the optimal value of (3)

Ideally we would want an unconstrained problem

$$\text{minimize } f_0(x) + \sum_{i=1}^{m} I_- \left(f_i(x) \right),$$

where $I_-(u) = \begin{cases} 0 & u \leq 0, \\ \infty & u > 0. \end{cases}$

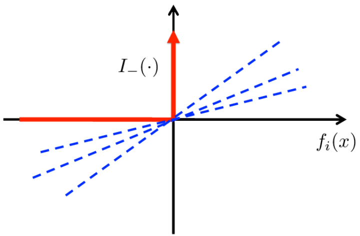
Why is this hard to solve?

Lower bound interpretation of Lagrangian

The **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is an (easier to optimize) lower bound on the original problem:

$$L(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \underbrace{\lambda_i f_i(x)}_{\leq I_-(f_i(x))},$$

The λ_i are called **lagrange multipliers** or **dual variables**.
To ensure a lower bound, we require $\lambda \succeq 0$.



Lower bound interpretation of Lagrangian

The **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is an (easier to optimize) lower bound on the original problem:

$$L(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \underbrace{\lambda_i f_i(x)}_{\leq I_-(f_i(x))},$$

The λ_i are called **lagrange multipliers** or **dual variables**.

Why bother?

- The original problem was very hard to solve (constraints). Minimizing the lower bound is easier (and can easily find the *closest* lower bound).
- Under "some conditions", the closest lower bound is tight: here minimum of $L(x, \lambda)$ at true x^* corresponding to p^*.

Lower bound interpretation of Lagrangian

The **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is an (easier to optimize) lower bound on the original problem:

$$L(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \underbrace{\lambda_i f_i(x)}_{\leq I_-(f_i(x))},$$

The λ_i are called **lagrange multipliers** or **dual variables**.

Why bother?

- The original problem was very hard to solve (constraints). Minimizing the lower bound is easier (and can easily find the *closest* lower bound).
- Under "some conditions", the closest lower bound is tight: here minimum of $L(x, \lambda)$ at true x^* corresponding to p^*.

Lagrange dual: lower bound on optimum p^*

The **Lagrange dual function:** minimize Lagrangian
When $\lambda \succeq 0$ and $f_i(x) \leq 0$, Lagrange dual function is

$$g(\lambda) := \inf_x L(x, \lambda). \tag{4}$$

We will show: (next slides) for any $\lambda \succeq 0$,

$$g(\lambda) \leq f_0(x)$$

wherever

$$\begin{aligned} f_i(x) &\leq 0 \\ g_i(x) &= 0 \end{aligned}$$

(including at $f_0(x^*) = p^*$).

Lagrange dual: lower bound on optimum p^*

The **Lagrange dual function:** minimize Lagrangian
When $\lambda \succeq 0$ and $f_i(x) \leq 0$, Lagrange dual function is

$$g(\lambda) := \inf_x L(x, \lambda). \tag{4}$$

We will show: (next slides) for any $\lambda \succeq 0$,

$$g(\lambda) \leq f_0(x)$$

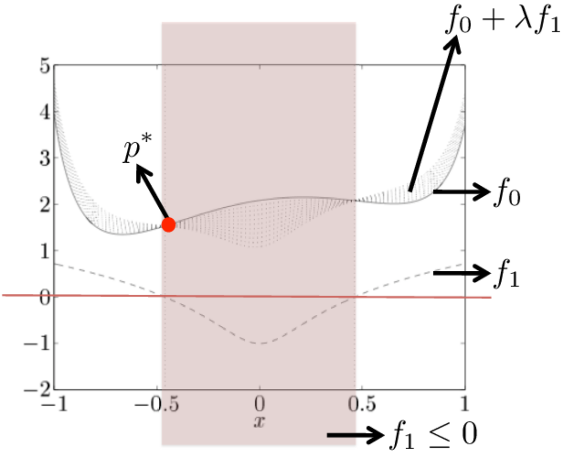
wherever

$$\begin{aligned} f_i(x) &\leq 0 \\ g_i(x) &= 0 \end{aligned}$$

(including at $f_0(x^*) = p^*$).

Lagrange dual: lower bound on optimum p^*

Simplest example: minimize over x the function $L(x, \lambda) = f_0(x) + \lambda f_1(x)$
(Figure from Boyd and Vandenberghe)



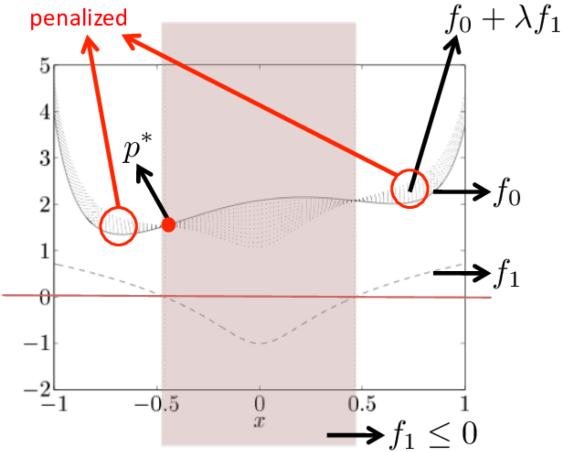
Reminders:

- f_0 is function to be minimized.
- $f_1 \leq 0$ is inequality constraint
- $\lambda \geq 0$ is Lagrange multiplier
- p^* is minimum f_0 *in constraint set*

Lagrange dual: lower bound on optimum p^*

Simplest example: minimize over x the function $L(x, \lambda) = f_0(x) + \lambda f_1(x)$

(Figure from Boyd and Vandenberghe)

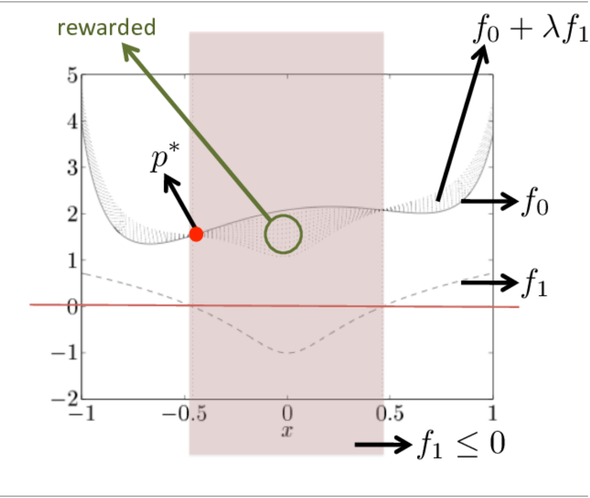


Reminders:

- f_0 is function to be minimized.
- $f_1 \leq 0$ is inequality constraint
- $\lambda \geq 0$ is Lagrange multiplier
- p^* is minimum f_0 *in constraint set*

Lagrange dual: lower bound on optimum p^*

Simplest example: minimize over x the function
$L(x, \lambda) = f_0(x) + \lambda f_1(x)$
(Figure from Boyd and Vandenberghe)



Reminders:

- f_0 is function to be minimized.
- $f_1 \leq 0$ is inequality constraint
- $\lambda \geq 0$ is Lagrange multiplier
- p^* is minimum f_0 *in constraint set*

Lagrange dual is lower bound on p^* (proof)

We now give a formal proof that **Lagrange dual function** $g(\lambda, \nu)$ lower bounds p^*.

Proof: Define $\tilde{x}$ as "some point" that is feasible, i.e. $f_i(\tilde{x}) \leq 0$, $\lambda \succeq 0$. Then

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \leq 0$$

Thus

$$\begin{aligned} g(\lambda) &:= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right) \\ &\leq f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \\ &\leq f_0(\tilde{x}). \end{aligned}$$

This holds for every feasible $\tilde{x}$, hence lower bound holds.

Lagrange dual is lower bound on p^* (proof)

We now give a formal proof that **Lagrange dual function** $g(\lambda, \nu)$ lower bounds p^*.

Proof: Define $\tilde{x}$ as "some point" that is feasible, i.e. $f_i(\tilde{x}) \leq 0$, $\lambda \succeq 0$. Then

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \leq 0$$

Thus

$$\begin{aligned} g(\lambda) &:= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right) \\ &\leq f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \\ &\leq f_0(\tilde{x}). \end{aligned}$$

This holds for every feasible $\tilde{x}$, hence lower bound holds.

Lagrange dual is lower bound on p^* (proof)

We now give a formal proof that **Lagrange dual function** $g(\lambda, \nu)$ lower bounds p^*.

Proof: Define $\tilde{x}$ as "some point" that is feasible, i.e. $f_i(\tilde{x}) \leq 0$, $\lambda \succeq 0$. Then

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \leq 0$$

Thus

$$\begin{aligned} g(\lambda) &:= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right) \\ &\leq f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \\ &\leq f_0(\tilde{x}). \end{aligned}$$

This holds for every feasible $\tilde{x}$, hence lower bound holds.

Best lower bound: maximize the dual

Closest (i.e. biggest) lower bound $g(\lambda)$ on the optimal solution p^* of original problem: **Lagrange dual problem**

$$\begin{array}{rl} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \succeq 0. \end{array} \quad (5)$$

Dual optimal: solutions λ^* maximizing dual, d^* is optimal value (**dual always easy to maximize**: next slide).
Weak duality always holds:

$$d^* \leq p^*.$$

...but what is the point of finding a **biggest lower bound** on a **minimization problem**?

Best lower bound: maximize the dual

Closest (i.e. biggest) lower bound $g(\lambda)$ on the optimal solution p^* of original problem: **Lagrange dual problem**

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{5}$$

Dual optimal: solutions λ^* maximizing dual, d^* is optimal value (**dual always easy to maximize**: next slide).
Weak duality always holds:

$$d^* \leq p^*.$$

...but what is the point of finding a **biggest lower bound** on a **minimization problem**?

Best lower bound: maximize the dual

Best (i.e. biggest) lower bound $g(\lambda)$ on the optimal solution p^* of original problem: **Lagrange dual problem**

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{6}$$

Dual optimal: solutions λ^* to the dual problem, d^* is optimal value (**dual always easy to maximize**: next slide).
Weak duality always holds:

$$d^* \leq p^*.$$

Strong duality: (does **not** always hold, conditions given later):

$$d^* = p^*.$$

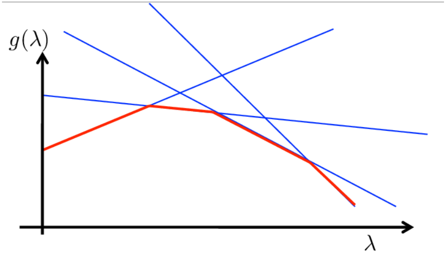
If S.D. holds: solve the **easy (concave) dual problem** to find p^*.

Maximizing the dual is always easy

The **Lagrange dual function:** minimize Lagrangian (lower bound)

$$g(\lambda) = \inf_{x \in \mathcal{D}} L(x, \lambda).$$

Dual function is a pointwise infimum of affine functions of λ, hence concave in λ with convex constraint set $\lambda \succeq 0$.



Example:

One inequality constraint,

$$L(x, \lambda) = f_0(x) + \lambda f_1(x),$$

and assume there are only four possible values for x. Each line represents a different x.

How do we know if strong duality holds?

Conditions under which strong duality holds are called **constraint qualifications** (they are sufficient, but not necessary)

(Probably) best known sufficient condition: Strong duality holds if

- Primal problem is **convex**, i.e. of the form

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0 \qquad\qquad i = 1, \ldots, n \end{aligned}$$

for convex f_0, affine $f_1, \ldots, f_m$.

A consequence of strong duality...

Assume primal is equal to the dual. What are the consequences?

- x^* solution of original problem (minimum of f_0 under constraints),
- λ^* solution to dual

$$
\begin{aligned}
f_0(x^*) &\underset{\text{(assumed)}}{=} g(\lambda^*) \\
&\underset{\text{(g definition)}}{=} \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) \right) \\
&\underset{\text{(inf definition)}}{\leq} f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) \\
&\underset{(4)}{\leq} f_0(x^*),
\end{aligned}
$$

(4): (x^*, λ^*) satisfies $\lambda^* \succeq 0$, and $f_i(x^*) \leq 0$.

...is complementary slackness

From previous slide,

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0, \tag{7}$$

which is the condition of **complementary slackness**. This means

$$\begin{aligned} \lambda_i^* > 0 \quad &\Longrightarrow \quad f_i(x^*) = 0, \\ f_i(x^*) < 0 \quad &\Longrightarrow \quad \lambda_i^* = 0. \end{aligned}$$

From λ_i, read off which inequality constraints are strict.

KKT conditions for global optimum

Assume functions f_i are differentiable and **strong duality**. Since x^* minimizes $L(x, \lambda^*)$, derivative at x^* is zero,

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) = 0.$$

KKT conditions definition: we are at **global optimum,** $(x, \lambda, \nu) = (x^*, \lambda^*)$ when (a) **strong duality** holds, and (b)

$$\begin{aligned} f_i(x) &\leq 0, \; i = 1, \ldots, m \\ \lambda_i &\geq 0, \; i = 1, \ldots, m \\ \lambda_i f_i(x) &= 0, \; i = 1, \ldots, m \\ \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) &= 0 \end{aligned}$$

KKT conditions for global optimum

Assume functions f_i are differentiable and **strong duality**. Since x^* minimizes $L(x, \lambda^*)$, derivative at x^* is zero,

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) = 0.$$

KKT conditions definition: we are at **global optimum**, $(x, \lambda, \nu) = (x^*, \lambda^*)$ when (a) **strong duality** holds, and (b)

$$\begin{aligned} f_i(x) &\leq 0, \; i = 1, \ldots, m \\ \lambda_i &\geq 0, \; i = 1, \ldots, m \\ \lambda_i f_i(x) &= 0, \; i = 1, \ldots, m \\ \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) &= 0 \end{aligned}$$

KKT conditions for global optimum

Assume functions f_i are differentiable and **strong duality**. Since x^* minimizes $L(x, \lambda^*)$, derivative at x^* is zero,

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) = 0.$$

KKT conditions definition: we are at **global optimum,** $(x, \lambda, \nu) = (x^*, \lambda^*)$ when (a) **strong duality** holds, and (b)

$$\begin{aligned} f_i(x) &\leq 0, \; i = 1, \ldots, m \\ \lambda_i &\geq 0, \; i = 1, \ldots, m \\ \lambda_i f_i(x) &= 0, \; i = 1, \ldots, m \\ \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) &= 0 \end{aligned}$$

KKT conditions for global optimum

In summary: if

- primal problem convex and
- inequality constraints affine

then strong duality holds. If in addition

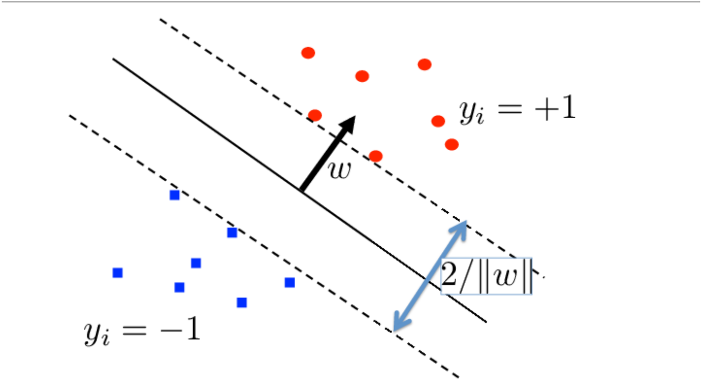
- functions f_i differentiable

then KKT conditions *necessary and sufficient* for optimality.

Support vector classification

Reminder: linearly separable points

Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



Smallest distance from each class to the separating hyperplane $w^\top x + b$ is called the **margin**.

Maximum margin classifier, linearly separable case

This problem can be expressed as follows:

$$\max_{w,b} (\text{margin}) = \max_{w,b} \left(\frac{2}{\|w\|} \right) \tag{8}$$

subject to

$$\begin{cases} w^\top x_i + b \geq 1 & i : y_i = +1, \\ w^\top x_i + b \leq -1 & i : y_i = -1. \end{cases} \tag{9}$$

The resulting classifier is

$$y = \text{sign}(w^\top x + b),$$

We can rewrite to obtain

$$\max_{w,b} \frac{1}{\|w\|} \quad \text{or} \quad \min_{w,b} \|w\|^2$$

subject to

$$y_i(w^\top x_i + b) \geq 1. \tag{10}$$

Maximum margin classifier, linearly separable case

This problem can be expressed as follows:

$$\max_{w,b}(\mathrm{margin}) = \max_{w,b}\left(\frac{2}{\|w\|}\right) \tag{8}$$

subject to

$$\begin{cases} w^\top x_i + b \geq 1 & i \ : \ y_i = +1, \\ w^\top x_i + b \leq -1 & i \ : \ y_i = -1. \end{cases} \tag{9}$$

The resulting classifier is

$$y = \mathrm{sign}(w^\top x + b),$$

We can rewrite to obtain

$$\max_{w,b} \frac{1}{\|w\|} \quad \mathrm{or} \quad \min_{w,b} \|w\|^2$$

subject to

$$y_i(w^\top x_i + b) \geq 1. \tag{10}$$

Maximum margin classifier: with errors allowed

Allow "errors": points within the margin, or even on the wrong side of the decision boundary. Ideally:

$$\min_{w,b}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\mathbb{I}[y_i\left(w^\top x_i + b\right) < 0]\right),$$

where C controls the tradeoff between maximum margin and loss.

...but this is too hard! (Why?)

Maximum margin classifier: with errors allowed

Allow "errors": points within the margin, or even on the wrong side of the decision boudary. Ideally:

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \mathbb{I}[y_i \left(w^\top x_i + b \right) < 0] \right),$$

where C controls the tradeoff between maximum margin and loss. Replace with **convex upper bound**:

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \theta \left(y_i \left(w^\top x_i + b \right) \right) \right).$$

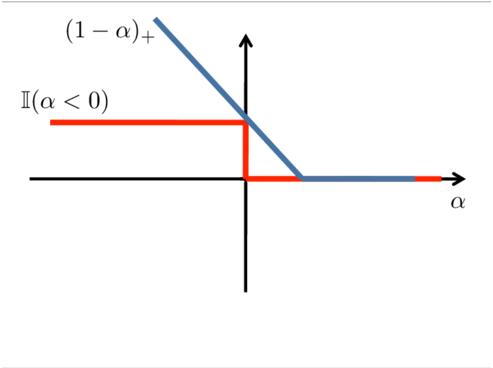
with hinge loss,

$$\theta(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha & 1-\alpha > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Hinge loss

Hinge loss:

$$\theta(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha & 1-\alpha > 0 \\ 0 & \text{otherwise.} \end{cases}$$



Support vector classification

Substituting in the hinge loss, we get

$$\min_{w,b}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\theta\left(y_i\left(w^\top x_i + b\right)\right)\right).$$

How do you implement hinge loss with simple **inequality constraints** (i.e. for convex optimization)?

$$\min_{w,b,\xi}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right) \tag{11}$$

subject to[1]

$$\xi_i \geq 0 \qquad y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i$$

[1] Either $y_i\left(w^\top x_i + b\right) \geq 1$ and $\xi_i = 0$ as before, or $y_i\left(w^\top x_i + b\right) < 1$, and then $\xi_i > 0$ takes the value satisfying $y_i\left(w^\top x_i + b\right) = 1 - \xi_i$.

Support vector classification

Substituting in the hinge loss, we get

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \theta \left(y_i \left(w^\top x_i + b \right) \right) \right).$$

How do you implement hinge loss with simple **inequality constraints** (i.e. for convex optimization)?

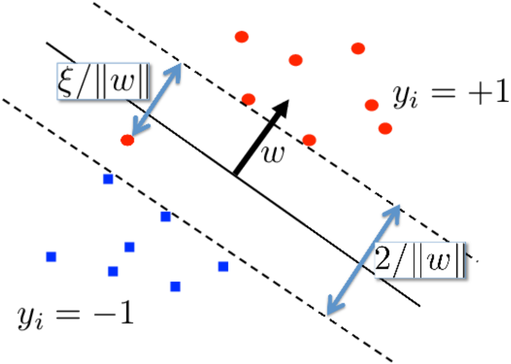
$$\min_{w,b,\xi} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \right) \tag{11}$$

subject to[1]

$$\xi_i \geq 0 \qquad y_i \left(w^\top x_i + b \right) \geq 1 - \xi_i$$

[1] Either $y_i \left(w^\top x_i + b \right) \geq 1$ and $\xi_i = 0$ as before, or $y_i \left(w^\top x_i + b \right) < 1$, and then $\xi_i > 0$ takes the value satisfying $y_i \left(w^\top x_i + b \right) = 1 - \xi_i$.

Support vector classification



Support vector classification

1. Convex optimization problem over the variables w, b, ξ:

$$\begin{aligned} \text{minimize} \quad & f_0(w,b,\xi) := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & 1 - \xi_i - y_i\left(w^\top x_i + b\right) \le 0 \qquad i = 1,\ldots,n \\ & -\xi_i \le 0 \qquad i = 1,\ldots,n \end{aligned}$$

(f_0 is convex, $f_1, \ldots, f_n$ are affine).

Strong duality holds, **and** the problem is differentiable, hence the KKT conditions hold at the global optimum.

Support vector classification

1. Convex optimization problem over the variables w, b, ξ:

$$\begin{aligned} \text{minimize} \quad & f_0(w,b,\xi) := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & 1 - \xi_i - y_i\left(w^\top x_i + b\right) \leq 0 \qquad i = 1,\ldots,n \\ & -\xi_i \leq 0 \qquad i = 1,\ldots,n \end{aligned}$$

(f_0 is convex, $f_1, \ldots, f_n$ are affine).

Strong duality holds, **and** the problem is differentiable, hence the KKT conditions hold at the global optimum.

Support vector classification

1. Convex optimization problem over the variables w, b, ξ:

$$\begin{aligned} \text{minimize} \quad & f_0(w,b,\xi) := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & 1-\xi_i - y_i\left(w^\top x_i + b\right) \le 0 \qquad i = 1,\ldots,n \\ & -\xi_i \le 0 \qquad i = 1,\ldots,n \end{aligned}$$

(f_0 is convex, $f_1, \ldots, f_n$ are affine).

Strong duality holds, **and** the problem is differentiable, hence the KKT conditions hold at the global optimum.

Support vector classification: Lagrangian

The Lagrangian: $L(w, b, \xi, \alpha, \lambda)$

$$= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i \left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) + \sum_{i=1}^n \lambda_i(-\xi_i)$$

with dual variable constraints

$$\alpha_i \geq 0, \qquad \lambda_i \geq 0.$$

Minimize wrt the primal variables w, b, and ξ.

Derivative wrt w:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i. \tag{12}$$

Derivative wrt b:

$$\frac{\partial L}{\partial b} = \sum_i y_i \alpha_i = 0. \tag{13}$$

Support vector classification: Lagrangian

The Lagrangian: $L(w, b, \xi, \alpha, \lambda)$

$$= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i \left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) + \sum_{i=1}^n \lambda_i(-\xi_i)$$

with dual variable constraints

$$\alpha_i \geq 0, \qquad \lambda_i \geq 0.$$

Minimize wrt the primal variables w, b, and ξ.

Derivative wrt w:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i. \tag{12}$$

Derivative wrt b:

$$\frac{\partial L}{\partial b} = \sum_i y_i \alpha_i = 0. \tag{13}$$

Support vector classification: Lagrangian

Derivative wrt ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \qquad \alpha_i = C - \lambda_i. \tag{14}$$

Noting that $\lambda_i \geq 0$,

$$\alpha_i \leq C.$$

Now use complementary slackness:

Non-margin SVs: $\alpha_i = C \neq 0$:

1. We immediately have $1 - \xi_i = y_i \left(w^\top x_i + b\right)$.
2. Also, from condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$, hence possibly $\xi_i > 0$.

Margin SVs: $0 < \alpha_i < C$:

1. We again have $1 - \xi_i = y_i \left(w^\top x_i + b\right)$
2. This time, from $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Non-SVs: $\alpha_i = 0$

1. We can allow: $y_i \left(w^\top x_i + b\right) > 1 - \xi_i$
2. From $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Support vector classification: Lagrangian

Derivative wrt ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \qquad \alpha_i = C - \lambda_i. \tag{14}$$

Noting that $\lambda_i \geq 0$,

$$\alpha_i \leq C.$$

Now use complementary slackness:

Non-margin SVs: $\alpha_i = C \neq 0$:

1. We immediately have $1 - \xi_i = y_i \left(w^\top x_i + b \right)$.
2. Also, from condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$, hence possibly $\xi_i > 0$.

Margin SVs: $0 < \alpha_i < C$:

1. We again have $1 - \xi_i = y_i \left(w^\top x_i + b \right)$
2. This time, from $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Non-SVs: $\alpha_i = 0$

1. We can allow: $y_i \left(w^\top x_i + b \right) > 1 - \xi_i$
2. From $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Support vector classification: Lagrangian

Derivative wrt ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \qquad \alpha_i = C - \lambda_i. \tag{14}$$

Noting that $\lambda_i \geq 0$,

$$\alpha_i \leq C.$$

Now use complementary slackness:

Non-margin SVs: $\alpha_i = C \neq 0$:

1. We immediately have $1 - \xi_i = y_i \left(w^\top x_i + b \right)$.
2. Also, from condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$, hence possibly $\xi_i > 0$.

Margin SVs: $0 < \alpha_i < C$:

1. We again have $1 - \xi_i = y_i \left(w^\top x_i + b \right)$
2. This time, from $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Non-SVs: $\alpha_i = 0$

1. We can allow: $y_i \left(w^\top x_i + b \right) > 1 - \xi_i$
2. From $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Support vector classification: Lagrangian

Derivative wrt ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \qquad \alpha_i = C - \lambda_i. \tag{14}$$

Noting that $\lambda_i \geq 0$,

$$\alpha_i \leq C.$$

Now use complementary slackness:

Non-margin SVs: $\alpha_i = C \neq 0$:

1. We immediately have $1 - \xi_i = y_i \left(w^\top x_i + b \right)$.
2. Also, from condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$, hence possibly $\xi_i > 0$.

Margin SVs: $0 < \alpha_i < C$:

1. We again have $1 - \xi_i = y_i \left(w^\top x_i + b \right)$
2. This time, from $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Non-SVs: $\alpha_i = 0$

1. We can allow: $y_i \left(w^\top x_i + b \right) > 1 - \xi_i$
2. From $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

The support vectors

We observe:

1. The solution is sparse: points which are not on the margin, or "margin errors", have $\alpha_i = 0$
2. The support vectors: only those points on the decision boundary, or which are margin errors, contribute.
3. Influence of the non-margin SVs is bounded, since their weight cannot exceed C.

Support vector classification: dual function

Thus, our goal is to maximize the dual,

$$
\begin{aligned}
g(\alpha, \lambda) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) \\
&\quad + \sum_{i=1}^{n}\lambda_i(-\xi_i) \\
&= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j \\
&\quad \underbrace{-b\sum_{i=1}^{m}\alpha_i y_i}_{0} + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}(C - \alpha_i)\xi_i \\
&= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j.
\end{aligned}
$$

Support vector classification: dual function

Maximize the dual,

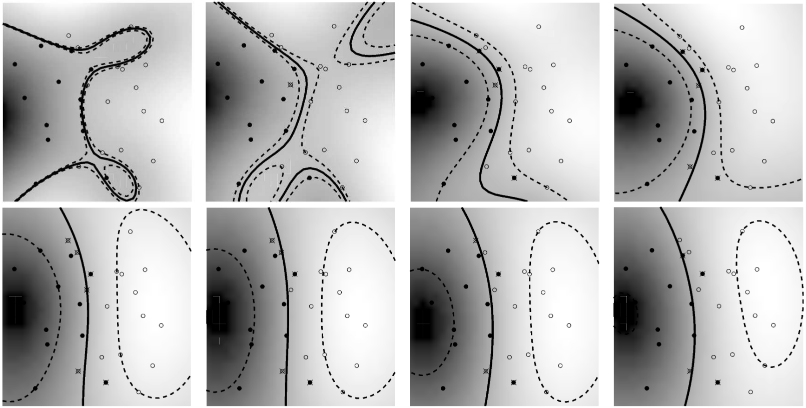
$$g(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^\top x_j,$$

subject to the constraints

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

This is a quadratic program.
Offset b: for the margin SVs, we have $1 = y_i \left(w^\top x_i + b \right)$. Obtain b from any of these, or take an average.

Support vector classification: kernel version



Taken from Schoelkopf and Smola (2002)

Support vector classification: kernel version

Maximum margin classifier in RKHS: write the hinge loss formulation

$$\min_{w} \left(\frac{1}{2} \|w(\cdot)\|_{\mathcal{H}}^2 + C \sum_{i=1}^{n} \theta \left(y_i \left\langle w(\cdot), \phi(x_i) \right\rangle_{\mathcal{H}} \right) \right)$$

for the RKHS $\mathcal{H}$ with kernel $k(x, \cdot)$. Use the result of the representer theorem,

$$w(\cdot) = \sum_{i=1}^{n} \beta_i \phi(x_i).$$

Maximizing the margin equivalent to minimizing $\|w(\cdot)\|_{\mathcal{H}}^2$: for many RKHSs a smoothness constraint (e.g. Gaussian kernel).

Support vector classification: kernel version

Substituting and introducing the ξ_i variables, get

$$\min_{\beta,\xi} \left(\frac{1}{2}\beta^\top K \beta + C \sum_{i=1}^{n} \xi_i \right) \tag{15}$$

where the matrix K has i,jth entry $K_{ij} = k(x_i, x_j)$, subject to

$$\xi_i \geq 0 \qquad y_i \sum_{j=1}^{n} \beta_j k(x_i, x_j) \geq 1 - \xi_i$$

Convex in β, ξ since K is positive definite.

Dual:

$$g(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j),$$

subject to the constraints $0 \leq \alpha_i \leq C$, and

$$w(\cdot) = \sum_{i=1}^{n} y_i \alpha_i \phi(x_i).$$

Support vector classification: kernel version

Substituting and introducing the ξ_i variables, get

$$\min_{\beta,\xi} \left(\frac{1}{2}\beta^\top K \beta + C \sum_{i=1}^{n} \xi_i \right) \tag{15}$$

where the matrix K has i,jth entry $K_{ij} = k(x_i, x_j)$, subject to

$$\xi_i \geq 0 \qquad y_i \sum_{j=1}^{n} \beta_j k(x_i, x_j) \geq 1 - \xi_i$$

Convex in β, ξ since K is positive definite.

Dual:

$$g(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j),$$

subject to the constraints $0 \leq \alpha_i \leq C$, and

$$w(\cdot) = \sum_{i=1}^{n} y_i \alpha_i \phi(x_i).$$

Questions?

Representer theorem

Learning problem: setting

Given a set of paired observations $(x_1, y_1), \ldots (x_n, y_n)$ (regression or classification).
Find the function f^* in the RKHS $\mathcal{H}$ which satisfies

$$J(f^*) = \min_{f \in \mathcal{H}} J(f), \tag{16}$$

where

$$J(f) = L_y(f(x_1), \ldots, f(x_n)) + \Omega\left(\|f\|^2_{\mathcal{H}}\right),$$

Ω is non-decreasing, and y is the vector of y_i.

- Classification: $L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^n \mathbb{I}_{y_i f(x_i) \leq 0}$
- Regression: $L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^n (y_i - f(x_i))^2$

Representer theorem

The representer theorem: solution to

$$\min_{f\in\mathcal{H}}\left[L_y(f(x_1),\ldots,f(x_n))+\Omega\left(\|f\|_{\mathcal{H}}^2\right)\right]$$

takes the form

$$f^*=\sum_{i=1}^n \alpha_i\phi(x_i).$$

If Ω is strictly increasing, all solutions have this form.

Representer theorem: proof

Proof: Denote f_s projection of f onto the subspace

$$\mathrm{span}\left\{\phi(x_i):\ 1 \leq i \leq n\right\}, \qquad (17)$$

such that

$$f = f_s + f_\perp,$$

where $f_s = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$.
Regularizer:

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2,$$

then

$$\Omega\left(\|f\|_{\mathcal{H}}^2\right) \geq \Omega\left(\|f_s\|_{\mathcal{H}}^2\right),$$

so this term is minimized for $f = f_s$.

Representer theorem: proof

Proof (cont.): Individual terms $f(x_i)$ in the loss:

$$f(x_i) = \langle f, \phi(x_i) \rangle_{\mathcal{H}} = \langle f_s + f_\perp, \phi(x_i) \rangle_{\mathcal{H}} = \langle f_s, \phi(x_i) \rangle_{\mathcal{H}},$$

so

$$L_y(f(x_1), \ldots, f(x_n)) = L_y(f_s(x_1), \ldots, f_s(x_n)).$$

Hence

- Loss $L(\ldots)$ only depends on the component of f in the data subspace,
- Regularizer $\Omega(\ldots)$ minimized when $f = f_s$.
- If Ω is strictly non-decreasing, then $\|f_\perp\|_{\mathcal{H}} = 0$ is required at the minimum.

Support vector classification: the ν-SVM

Hard to interpret C. Modify the formulation to get a more intuitive parameter ν.

Again, we drop b for simplicity. Solve

$$\min_{w,\rho,\xi}\left(\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i\right)$$

subject to

$$\begin{aligned} \rho &\geq 0 \\ \xi_i &\geq 0 \\ y_i w^\top x_i &\geq \rho - \xi_i, \end{aligned}$$

(now directly adjust margin width ρ).

The ν-SVM: Lagrangian

$$\frac{1}{2}\|w\|^2+\frac{1}{n}\sum_{i=1}^{n}\xi_i-\nu\rho+\sum_{i=1}^{n}\alpha_i\left(\rho-y_iw^\top x_i-\xi_i\right)+\sum_{i=1}^{n}\beta_i(-\xi_i)+\gamma(-\rho)$$

for dual variables $\alpha_i \geq 0$, $\beta_i \geq 0$, and $\gamma \geq 0$.

Differentiating and setting to zero for each of the primal variables w, ξ, ρ,

$$w = \sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\alpha_i + \beta_i = \frac{1}{n} \quad (18)$$

$$\nu = \sum_{i=1}^{n}\alpha_i - \gamma \quad (19)$$

From $\beta_i \geq 0$, equation (18) implies

$$0 \leq \alpha_i \leq n^{-1}.$$

The ν-SVM: Lagrangian

$$\frac{1}{2}\|w\|^2+\frac{1}{n}\sum_{i=1}^n \xi_i-\nu\rho+\sum_{i=1}^n \alpha_i\left(\rho-y_i w^\top x_i-\xi_i\right)+\sum_{i=1}^n \beta_i(-\xi_i)+\gamma(-\rho)$$

for dual variables $\alpha_i \geq 0$, $\beta_i \geq 0$, and $\gamma \geq 0$.
Differentiating and setting to zero for each of the primal variables w, ξ, ρ,

$$\begin{aligned} w &= \sum_{i=1}^n \alpha_i y_i x_i \\ \alpha_i+\beta_i &= \frac{1}{n} \quad (18) \\ \nu &= \sum_{i=1}^n \alpha_i-\gamma \quad (19) \end{aligned}$$

From $\beta_i \geq 0$, equation (18) implies

$$0 \leq \alpha_i \leq n^{-1}.$$

Complementary slackness (1)

Complementary slackness conditions:

Assume $\rho > 0$ at the global solution, hence $\gamma = 0$, and

$$\sum_{i=1}^{n} \alpha_i = \nu. \tag{20}$$

Case of $\xi_i > 0$: complementary slackness states $\beta_i = 0$, hence from (18) we have $\alpha_i = n^{-1}$. Denote this set as $N(\alpha)$. Then

$$\sum_{i \in N(\alpha)} \frac{1}{n} = \sum_{i \in N(\alpha)} \alpha_i \leq \sum_{i=1}^{n} \alpha_i = \nu,$$

so

$$\frac{|N(\alpha)|}{n} \leq \nu,$$

and ν is an upper bound on the number of non-margin SVs.

Complementary slackness conditions:

Assume $\rho > 0$ at the global solution, hence $\gamma = 0$, and

$$\sum_{i=1}^{n} \alpha_i = \nu. \qquad (20)$$

Case of $\xi_i > 0$: complementary slackness states $\beta_i = 0$, hence from (18) we have $\alpha_i = n^{-1}$. Denote this set as $N(\alpha)$. Then

$$\sum_{i \in N(\alpha)} \frac{1}{n} = \sum_{i \in N(\alpha)} \alpha_i \leq \sum_{i=1}^{n} \alpha_i = \nu,$$

so

$$\frac{|N(\alpha)|}{n} \leq \nu,$$

and ν is an upper bound on the number of non-margin SVs.

Complementary slackness (2)

Case of $\xi_i = 0$: $\alpha_i < n^{-1}$. Denote by $M(\alpha)$ the set of points $n^{-1} > \alpha_i > 0$. Then from (20),

$$\nu = \sum_{i=1}^{n} \alpha_i = \sum_{i \in N(\alpha)} \frac{1}{n} + \sum_{i \in M(\alpha)} \alpha_i \leq \sum_{i \in M(\alpha) \cup N(\alpha)} \frac{1}{n},$$

thus

$$\nu \leq \frac{|N(\alpha)| + |M(\alpha)|}{n},$$

and ν is a lower bound on the number of support vectors with non-zero weight (both on the margin, and "margin errors").

Dual for ν-SVM

Substituting into the Lagrangian, we get

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \rho\nu - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j$$

$$+\sum_{i=1}^{n}\alpha_i\rho - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\left(\frac{1}{n}-\alpha_i\right)\xi_i - \rho\left(\sum_{i=1}^{n}\alpha_i - \nu\right)$$

$$= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j$$

Maximize:

$$g(\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j,$$

subject to

$$\sum_{i=1}^{n}\alpha_i \geq \nu \qquad 0 \leq \alpha_i \leq \frac{1}{n}.$$

Dual for ν-SVM

Substituting into the Lagrangian, we get

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \rho\nu - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j$$
$$+\sum_{i=1}^{n}\alpha_i\rho - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\left(\frac{1}{n}-\alpha_i\right)\xi_i - \rho\left(\sum_{i=1}^{n}\alpha_i - \nu\right)$$
$$= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j$$

Maximize:

$$g(\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j,$$

subject to

$$\sum_{i=1}^{n}\alpha_i \geq \nu \qquad 0 \leq \alpha_i \leq \frac{1}{n}.$$