Lecture 2: Linear SVM in the Dual

Stéphane Canu
stephane.canu@litislab.eu

Sao Paulo 2014

March 12, 2014

Road map

1. Linear SVM
   - Optimization in 10 slides
     - Equality constraints
     - Inequality constraints
   - Dual formulation of the linear SVM
   - Solving the dual

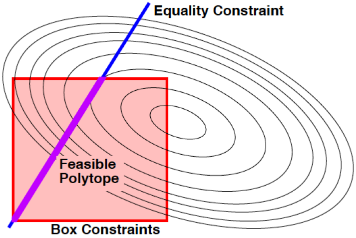


Figure from L. Bottou & C.J. Lin, Support vector machine solvers, in Large scale kernel machines, 2007.

Linear SVM: the problem

Linear SVM are the solution of the following problem (called primal)

Let $\{(\mathbf{x}_i, y_i);\ i = 1:n\}$ be a set of labelled data with $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{1, -1\}$.
A support vector machine (SVM) is a linear classifier associated with the following decision function: $D(x) = \text{sign}\big(\mathbf{w}^\top \mathbf{x} + b\big)$ where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ a given thought the solution of the following problem:

$$\begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w}^\top\mathbf{w} \\ \text{with} & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 \qquad\qquad i = 1, n \end{cases}$$

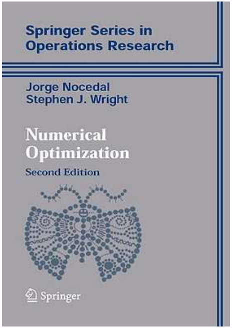
This is a quadratic program (QP):

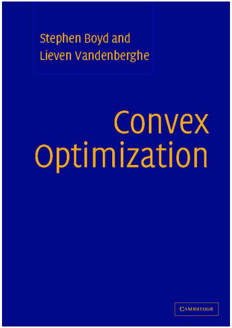
$$\begin{cases} \min\limits_{\mathbf{z}} & \frac{1}{2}\mathbf{z}^\top A\mathbf{z} - \mathbf{d}^\top\mathbf{z} \\ \text{with} & B\mathbf{z} \leq \mathbf{e} \end{cases}$$

$\mathbf{z} = (\mathbf{w}, b)^\top$, $\mathbf{d} = (0, \ldots, 0)^\top$, $A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$, $B = -[\text{diag}(\mathbf{y})X, \mathbf{y}]$ et $\mathbf{e} = -(1, \ldots, 1)^\top$

Road map

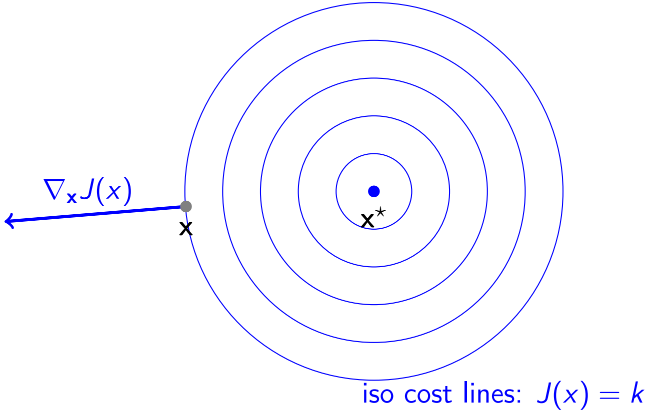
1. Linear SVM
   - Optimization in 10 slides
     - Equality constraints
     - Inequality constraints
   - Dual formulation of the linear SVM
   - Solving the dual





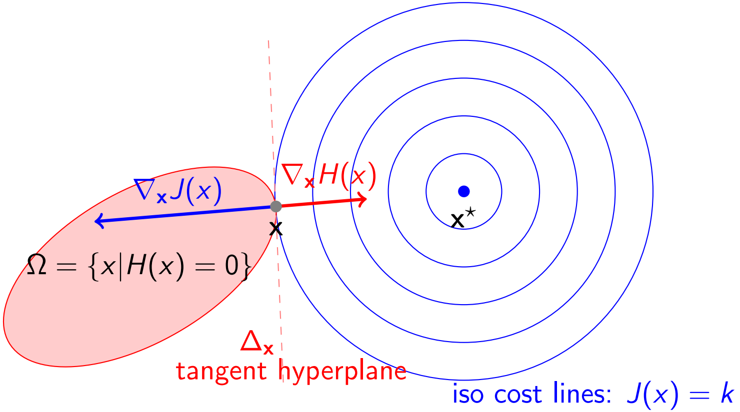
A simple example (to begin with)

$$\begin{cases} \min\limits_{x_1, x_2} & J(\mathbf{x}) = (x_1 - a)^2 + (x_2 - b)^2 \\ \text{with} & \end{cases}$$



A simple example (to begin with)

$$\begin{cases} \min\limits_{x_1,x_2} & J(\mathbf{x}) = (x_1 - a)^2 + (x_2 - b)^2 \\ \text{with} & H(\mathbf{x}) = \alpha(x_1 - c)^2 + \beta(x_2 - d)^2 + \gamma x_1 x_2 - 1 \end{cases}$$



$$\nabla_{\mathbf{x}} H(x) = \lambda \, \nabla_{\mathbf{x}} J(x)$$

The only one equality constraint case

$$\begin{cases} \min\limits_{\mathbf{x}} & J(\mathbf{x}) \\ \text{with} & H(\mathbf{x}) = 0 \end{cases} \qquad \begin{aligned} J(\mathbf{x}+\varepsilon\mathbf{d}) &\approx J(\mathbf{x}) + \varepsilon\nabla_{\mathbf{x}}J(\mathbf{x})^\top\mathbf{d} \\ H(\mathbf{x}+\varepsilon\mathbf{d}) &\approx H(\mathbf{x}) + \varepsilon\nabla_{\mathbf{x}}H(\mathbf{x})^\top\mathbf{d} \end{aligned}$$

Loss J : $\mathbf{d}$ is a descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall\varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$J(\mathbf{x}+\varepsilon\mathbf{d}) < J(\mathbf{x}) \qquad \Rightarrow \qquad \nabla_{\mathbf{x}}J(\mathbf{x})^\top\mathbf{d} < 0$$

constraint H : $\mathbf{d}$ is a feasible descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall\varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$H(\mathbf{x}+\varepsilon\mathbf{d}) = 0 \qquad \Rightarrow \qquad \nabla_{\mathbf{x}}H(\mathbf{x})^\top\mathbf{d} = 0$$

If at $x^\star$, vectors $\nabla_{\mathbf{x}}J(\mathbf{x}^\star)$ and $\nabla_{\mathbf{x}}H(\mathbf{x}^\star)$ are collinear there is no feasible descent direction $\mathbf{d}$. Therefore, $x^\star$ is a local solution of the problem.

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) \\ \text{avec} & H_1(\mathbf{x}) = 0 \\ \text{et} & H_2(\mathbf{x}) = 0 \\ & \dots \\ & H_p(\mathbf{x}) = 0 \end{cases}$$

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \displaystyle\min_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) \\ \text{avec} & H_1(\mathbf{x}) = 0 \qquad\qquad \lambda_1 \\ \text{et} & H_2(\mathbf{x}) = 0 \qquad\qquad \lambda_2 \\ & \dots \\ & H_p(\mathbf{x}) = 0 \qquad\qquad \lambda_p \end{cases}$$

each constraint is associated with λ_i : the Lagrange multiplier.

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) & \\ \text{avec} & H_1(\mathbf{x}) = 0 & \qquad\qquad {\color{red}\lambda_1} \\ \text{et} & H_2(\mathbf{x}) = 0 & \qquad\qquad {\color{red}\lambda_2} \\ & \dots & \\ & H_p(\mathbf{x}) = 0 & \qquad\qquad {\color{red}\lambda_p} \end{cases}$$

each constraint is associated with λ_i : the Lagrange multiplier.

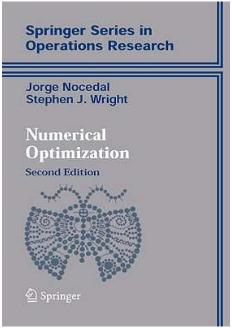
Theorem (First order optimality conditions)

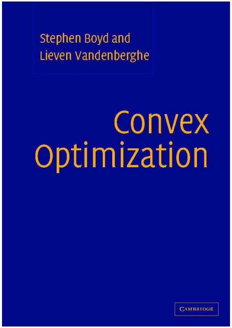
for $\mathbf{x}^\star$ being a local minima of $\mathcal{P}$, it is necessary that:

$$\nabla_x J(\mathbf{x}^\star) + \sum_{i=1}^{p} \lambda_i \nabla_x H_i(\mathbf{x}^\star) = 0 \qquad \textit{and} \qquad H_i(\mathbf{x}^\star) = 0, \quad i = 1, p$$

Plan

1. Linear SVM
   - Optimization in 10 slides
     - Equality constraints
     - Inequality constraints
   - Dual formulation of the linear SVM
   - Solving the dual





The only one inequality constraint case

$$\begin{cases} \min_{\mathbf{x}} & J(\mathbf{x}) \\ \text{with} & G(\mathbf{x}) \leq 0 \end{cases} \qquad \begin{aligned} J(\mathbf{x}+\varepsilon\mathbf{d}) &\approx J(\mathbf{x}) + \varepsilon \nabla_{\mathbf{x}} J(\mathbf{x})^\top \mathbf{d} \\ G(\mathbf{x}+\varepsilon\mathbf{d}) &\approx G(\mathbf{x}) + \varepsilon \nabla_{\mathbf{x}} G(\mathbf{x})^\top \mathbf{d} \end{aligned}$$

cost J : **d** is a descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$J(\mathbf{x}+\varepsilon\mathbf{d}) < J(\mathbf{x}) \qquad \Rightarrow \qquad \nabla_{\mathbf{x}} J(\mathbf{x})^\top \mathbf{d} < 0$$

constraint G : **d** is a feasible descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$G(\mathbf{x}+\varepsilon\mathbf{d}) \leq 0 \qquad \Rightarrow \qquad \begin{array}{ll} G(\mathbf{x}) < 0: & \text{no limit here on } \mathbf{d} \\ G(\mathbf{x}) = 0: & \nabla_{\mathbf{x}} G(\mathbf{x})^\top \mathbf{d} \leq 0 \end{array}$$

Two possibilities

If $x^\star$ lies at the limit of the feasible domain ($G(\mathbf{x}^\star) = 0$) and if vectors $\nabla_{\mathbf{x}} J(\mathbf{x}^\star)$ and $\nabla_{\mathbf{x}} G(\mathbf{x}^\star)$ are collinear and in opposite directions, there is no feasible descent direction **d** at that point. Therefore, $x^\star$ is a local solution of the problem... Or if $\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = 0$

Two possibilities for optimality

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = -\mu \, \nabla_{\mathbf{x}} G(\mathbf{x}^\star) \qquad \text{and} \qquad \mu > 0;\, G(\mathbf{x}^\star) = 0$$

or

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = 0 \qquad \text{and} \qquad \mu = 0;\, G(\mathbf{x}^\star) < 0$$

This alternative is summarized in the so called complementarity condition:

$$\mu \;\; G(x^\star) = 0$$

$\mu = 0$
$G(x^\star) < 0$

$G(x^\star) = 0$
$\mu > 0$

First order optimality condition (1)

$$\text{problem } \mathcal{P} = \begin{cases} \min_{\mathbf{x}\in\mathbb{R}^n} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, \dots, p \\ \text{and} & g_i(x) \le 0 \;\; i = 1, \dots, q \end{cases}$$

Definition: Karush, Kuhn and Tucker (KKT) conditions

$$\begin{array}{rll} \text{stationarity} & \nabla J(x^\star) + \sum_{j=1}^{p} \lambda_j \nabla h_j(x^\star) + \sum_{i=1}^{q} \mu_i \nabla g_i(x^\star) = 0 & \\ \text{primal admissibility} & h_j(x^\star) = 0 & j = 1, \dots, p \\ & g_i(x^\star) \le 0 & i = 1, \dots, q \\ \text{dual admissibility} & \mu_i \ge 0 & i = 1, \dots, q \\ \text{complementarity} & \mu_i g_i(x^\star) = 0 & i = 1, \dots, q \end{array}$$

λ_j and μ_i are called the Lagrange multipliers of problem $\mathcal{P}$

First order optimality condition (2)

Theorem (12.1 Nocedal & Wright pp 321)

If a vector $x^\star$ is a stationary point of problem $\mathcal{P}$
Then there exists[a] Lagrange multipliers such that $\big(x^\star, \{\lambda_j\}_{j=1:p}, \{\mu_i\}_{i=1:q}\big)$ fulfill KKT conditions

[a] under some conditions *e.g.* linear independence constraint qualification

If the problem is convex, then a stationary point is the solution of the problem

A quadratic program (QP) is convex when. . .

$$(QP) \quad \begin{cases} \min\limits_{\mathbf{z}} & \frac{1}{2}\mathbf{z}^\top A\mathbf{z} - \mathbf{d}^\top \mathbf{z} \\ \text{with} & B\mathbf{z} \le \mathbf{e} \end{cases}$$

. . . when matrix A is positive definite

KKT condition - Lagrangian (3)

$$\text{problem } \mathcal{P} = \begin{cases} \min\limits_{\mathbf{x} \in \mathbb{R}^n} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, \dots, p \\ \text{and} & g_i(x) \leq 0 \; i = 1, \dots, q \end{cases}$$

Definition: Lagrangian

The lagrangian of problem $\mathcal{P}$ is the following function:

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = J(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x)$$

The importance of being a lagrangian

- the stationarity condition can be written: $\nabla \mathcal{L}(\mathbf{x}^\star, \lambda, \mu) = 0$
- the lagrangian saddle point $\quad \max\limits_{\lambda, \mu} \min\limits_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \mu)$

Primal variables: x and dual variables λ, μ (the Lagrange multipliers)

Duality – definitions (1)

Primal and (Lagrange) dual problems

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{x} \in \mathbb{R}^n} & J(\mathbf{x}) \\ \text{with} & h_j(x) = 0 \quad j = 1, p \\ \text{and} & g_i(x) \leq 0 \quad i = 1, q \end{cases} \qquad \mathcal{D} = \begin{cases} \max\limits_{\lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^q} & Q(\lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \end{cases}$$

Dual objective function:

$$\begin{aligned} Q(\lambda, \mu) &= \inf_x \; \mathcal{L}(\mathbf{x}, \lambda, \mu) \\ &= \inf_x \; J(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x) \end{aligned}$$

Wolf dual problem

$$\mathcal{W} = \begin{cases} \max\limits_{\mathbf{x}, \lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^q} & \mathcal{L}(\mathbf{x}, \lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \\ \text{and} & \nabla J(x^\star) + \sum_{j=1}^{p} \lambda_j \nabla h_j(x^\star) + \sum_{i=1}^{q} \mu_i \nabla g_i(x^\star) = 0 \end{cases}$$

Duality – theorems (2)

Theorem (12.12, 12.13 and 12.14 Nocedal & Wright pp 346)

If f, g and h are convex and continuously differentiable[a], then the solution of the dual problem is the same as the solution of the primal

[a] under some conditions *e.g.* linear independence constraint qualification

$$\begin{aligned}(\lambda^\star, \mu^\star) &= \text{solution of problem } \mathcal{D} \\ \mathbf{x}^\star &= \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^\star, \mu^\star)\end{aligned}$$

$$\begin{aligned}Q(\lambda^\star, \mu^\star) = \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^\star, \mu^\star) &= \mathcal{L}(\mathbf{x}^\star, \lambda^\star, \mu^\star) \\ &= J(\mathbf{x}^\star) + \lambda^\star H(\mathbf{x}^\star) + \mu^\star G(\mathbf{x}^\star) = J(\mathbf{x}^\star)\end{aligned}$$

and for any feasible point $\mathbf{x}$

$$Q(\lambda, \mu) \leq J(\mathbf{x}) \qquad \rightarrow \qquad 0 \leq J(\mathbf{x}) - Q(\lambda, \mu)$$

The duality gap is the difference between the primal and dual cost functions

Road map

1. Linear SVM
   - Optimization in 10 slides
     - Equality constraints
     - Inequality constraints
   - Dual formulation of the linear SVM
   - Solving the dual

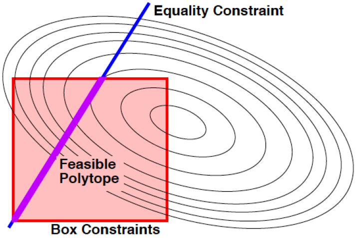


Figure from L. Bottou & C.J. Lin, Support vector machine solvers, in Large scale kernel machines, 2007.

Linear SVM dual formulation - The lagrangian

$$\begin{cases} \min\limits_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \qquad i = 1, n \end{cases}$$

Looking for the lagrangian saddle point $\max\limits_{\alpha} \min\limits_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha)$ with so called lagrange multipliers $\alpha_i \geq 0$

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\big)$$

α_i represents the influence of constraint thus the influence of the training example (x_i, y_i)

Stationarity conditions

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\big)$$

Computing the gradients: $\begin{cases} \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \alpha) & = \mathbf{w} - \sum\limits_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = \sum_{i=1}^{n} \alpha_i \; y_i \end{cases}$

we have the following optimality conditions

$$\begin{cases} \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \alpha) & = 0 \quad \Rightarrow \quad \mathbf{w} = \sum\limits_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \dfrac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} & = 0 \quad \Rightarrow \quad \sum\limits_{i=1}^{n} \alpha_i \; y_i = 0 \end{cases}$$

KKT conditions for SVM

stationarity	$\mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$ and $\sum_{i=1}^{n} \alpha_i \ y_i = 0$	
primal admissibility	$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$	$i = 1, \ldots, n$
dual admissibility	$\alpha_i \geq 0$	$i = 1, \ldots, n$
complementarity	$\alpha_i \Big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \Big) = 0$	$i = 1, \ldots, n$

The complementary condition split the data into two sets

- $\mathcal{A}$ be the set of active constraints: usefull points

$$\mathcal{A} = \{ i \in [1, n] \mid y_i(\mathbf{w}^{*\top} \mathbf{x}_i + b^*) = 1 \}$$

- its complementary $\bar{\mathcal{A}}$ useless points

$$\text{if } i \notin \mathcal{A}, \alpha_i = 0$$

The KKT conditions for SVM

The same KKT but using matrix notations and the active set $\mathcal{A}$

stationarity	$\boxed{\mathbf{w} - X^\top D_y \alpha = 0}$
	$\boxed{\alpha^\top y = 0}$
primal admissibility	$D_y(Xw + b\mathbb{1}) \geq \mathbb{1}$
dual admissibility	$\alpha \geq 0$
complementarity	$\boxed{D_y(X_{\mathcal{A}}\mathbf{w} + b\mathbb{1}_{\mathcal{A}}) = \mathbb{1}_{\mathcal{A}}}$
	$\alpha_{\bar{\mathcal{A}}} = 0$

Knowing $\mathcal{A}$, the solution verifies the following linear system:

$$\left\{ \begin{array}{cccl} \mathbf{w} & -X_{\mathcal{A}}^\top D_y \alpha_{\mathcal{A}} & & = 0 \\ -D_y X_{\mathcal{A}} \mathbf{w} & & -b\mathbf{y}_{\mathcal{A}} & = -\mathbf{e}_{\mathcal{A}} \\ & -\mathbf{y}_{\mathcal{A}}^\top \alpha_{\mathcal{A}} & & = 0 \end{array} \right.$$

with $D_y = \text{diag}(\mathbf{y}_{\mathcal{A}})$, $\alpha_{\mathcal{A}} = \alpha(\mathcal{A})$, $\mathbf{y}_{\mathcal{A}} = \mathbf{y}(\mathcal{A})$ et $X_{\mathcal{A}} = X(X_{\mathcal{A}};:)$.

The KKT conditions as a linear system

$$\left\{\begin{array}{cccl} \mathbf{w} & -X_{\mathcal{A}}^\top D_y \alpha_{\mathcal{A}} & & = 0 \\ -D_y X_{\mathcal{A}} \mathbf{w} & & -b\mathbf{y}_{\mathcal{A}} & = -\mathbf{e}_{\mathcal{A}} \\ & -\mathbf{y}_{\mathcal{A}}^\top \alpha_{\mathcal{A}} & & = 0 \end{array}\right.$$

with $D_y = \text{diag}(\mathbf{y}_{\mathcal{A}})$, $\alpha_{\mathcal{A}} = \alpha(\mathcal{A})$, $\mathbf{y}_{\mathcal{A}} = \mathbf{y}(\mathcal{A})$ et $X_{\mathcal{A}} = X(X_{\mathcal{A}};:)$.

$$\begin{bmatrix} I & -X_{\mathcal{A}}^\top D_y & 0 \\ -D_y X_{\mathcal{A}} & 0 & -\mathbf{y}_{\mathcal{A}} \\ 0 & -\mathbf{y}_{\mathcal{A}}^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \alpha_{\mathcal{A}} \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ -\mathbf{e}_{\mathcal{A}} \\ 0 \end{bmatrix}$$

we can work on it to separate **w** from $(\alpha_{\mathcal{A}}, b)$

The SVM dual formulation

The SVM Wolfe dual

$$\begin{cases} \max\limits_{\mathbf{w},b,\alpha} & \frac{1}{2}\|\mathbf{w}\|^2 - \sum\limits_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\big) \\ \text{with} & \alpha_i \geq 0 \qquad\qquad\qquad\qquad i = 1, \dots, n \\ \text{and} & \mathbf{w} - \sum\limits_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \text{ and } \sum\limits_{i=1}^{n} \alpha_i \; y_i = 0 \end{cases}$$

using the fact: $\mathbf{w} = \sum\limits_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$

The SVM Wolfe dual without $\mathbf{w}$ and b

$$\begin{cases} \max\limits_{\alpha} & -\frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i + \sum\limits_{i=1}^{n} \alpha_i \\ \text{with} & \alpha_i \geq 0 \qquad\qquad\qquad\qquad i = 1, \dots, n \\ \text{and} & \sum\limits_{i=1}^{n} \alpha_i \; y_i = 0 \end{cases}$$

Linear SVM dual formulation

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\big)$$

Optimality: $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad \sum_{i=1}^{n} \alpha_i \; y_i = 0$

$$\begin{aligned}
\mathcal{L}(\alpha) &= \underbrace{\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i}_{\mathbf{w}^\top \mathbf{w}} - \textstyle\sum_{i=1}^{n} \alpha_i y_i \underbrace{\sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j^\top}_{\mathbf{w}^\top} \mathbf{x}_i - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0} + \textstyle\sum_{i=1}^{n} \alpha_i \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j \alpha_i y_i y_j \mathbf{x}_j^\top \mathbf{x}_i + \sum_{i=1}^{n} \alpha_i
\end{aligned}$$

Dual linear SVM is also a quadratic program

$$\text{problem } \mathcal{D} \quad \begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G \alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \le \alpha_i \qquad\qquad i = 1, n \end{cases}$$

with G a symmetric matrix $n \times n$ such that $G_{ij} = y_i y_j \mathbf{x}_j^\top \mathbf{x}_i$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d + 1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d+1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i\, y_i(\mathbf{x}^\top\mathbf{x}_i) + b$$

The bi dual (the dual of the dual)

$$\begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

$$\begin{aligned} \mathcal{L}(\alpha, \lambda, \mu) &= \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha + \lambda\, \mathbf{y}^\top \alpha - \mu^\top \alpha \\ \nabla_\alpha \mathcal{L}(\alpha, \lambda, \mu) &= G\alpha - \mathbf{e} + \lambda\, \mathbf{y} - \mu \end{aligned}$$

The bidual

$$\begin{cases} \max\limits_{\alpha, \lambda, \mu} & -\frac{1}{2}\alpha^\top G\alpha \\ \text{with} & G\alpha - \mathbf{e} + \lambda\, \mathbf{y} - \mu = 0 \\ \text{and} & 0 \leq \mu \end{cases}$$

since $\|\mathbf{w}\|^2 = \frac{1}{2}\alpha^\top G\alpha$ and $DX\mathbf{w} = G\alpha$

$$\begin{cases} \max\limits_{\mathbf{w}, \lambda} & -\frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & DX\mathbf{w} + \lambda\, \mathbf{y} \geq \mathbf{e} \end{cases}$$

by identification (possibly up to a sign)

$b = \lambda$ is the Lagrange multiplier of the equality constraint

Cold case: the least square problem

Linear model

$$y_i = \sum_{j=1}^{d} \mathbf{w}_j x_{ij} + \varepsilon_i \quad , \qquad i = 1, n$$

n data and d variables; $d < n$

$$\min_{\mathbf{w}} = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} x_{ij} \mathbf{w}_j - y_i \right)^2 = \|X\mathbf{w} - \mathbf{y}\|^2$$

Solution: $\widetilde{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{y}$

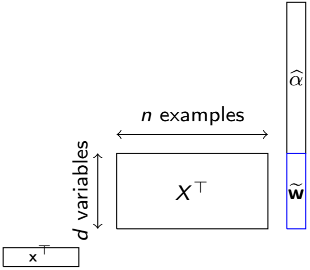
$$f(\mathbf{x}) = \mathbf{x}^\top \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{\mathbf{w}}}$$

What is the influence of each data point (matrix X lines) ?

data point influence (contribution)

for any new data point $\mathbf{x}$

$$
\begin{aligned}
f(\mathbf{x}) &= \mathbf{x}^\top (X^\top X)(X^\top X)^{-1} \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{\mathbf{w}}} \\
&= \mathbf{x}^\top X^\top \underbrace{X(X^\top X)^{-1}(X^\top X)^{-1} X^\top \mathbf{y}}_{\widehat{\alpha}}
\end{aligned}
$$



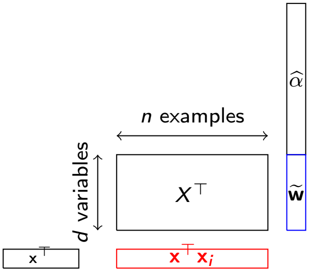
$$
f(\mathbf{x}) = \sum_{j=1}^{d} \widetilde{\mathbf{w}}_j x_j
$$

data point influence (contribution)

for any new data point $\mathbf{x}$

$$f(\mathbf{x}) = \mathbf{x}^\top (X^\top X)(X^\top X)^{-1} \underbrace{(X^\top X)^{-1} X^\top \mathbf{y}}_{\widetilde{\mathbf{w}}}$$

$$= \mathbf{x}^\top X^\top \underbrace{X(X^\top X)^{-1}(X^\top X)^{-1} X^\top \mathbf{y}}_{\widehat{\alpha}}$$



$$f(\mathbf{x}) = \sum_{j=1}^{d} \widetilde{\mathbf{w}}_j x_j = \sum_{i=1}^{n} \widehat{\alpha}_i \, (\mathbf{x}^\top \mathbf{x}_i)$$

from variables to examples

$$\underbrace{\widehat{\alpha} = X(X^\top X)^{-1} \widetilde{\mathbf{w}}}_{n \text{ examples}} \qquad \text{et} \qquad \underbrace{\widetilde{\mathbf{w}} = X^\top \widehat{\alpha}}_{d \text{ variables}}$$

what if $d \geq n$!

SVM primal vs. dual

Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{with} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \\ & i = 1, n \end{cases}$$

- $d+1$ unknown
- n constraints
- classical QP
- perfect when $d << n$

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha \\ \text{with} & \mathbf{y}^\top \alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases}$$

- n unknown
- G Gram matrix (pairwise influence matrix)
- n box constraints
- easy to solve
- to be used when $d > n$

$$f(\mathbf{x}) \;=\; \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \; y_i(\mathbf{x}^\top \mathbf{x}_i) + b$$

Road map

1. Linear SVM
   - Optimization in 10 slides
     - Equality constraints
     - Inequality constraints
   - Dual formulation of the linear SVM
   - Solving the dual

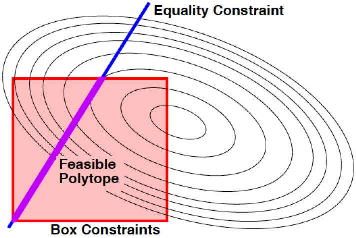
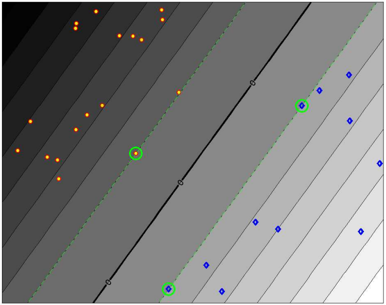


Figure from L. Bottou & C.J. Lin, Support vector machine solvers, in Large scale kernel machines, 2007.

Solving the dual (1)

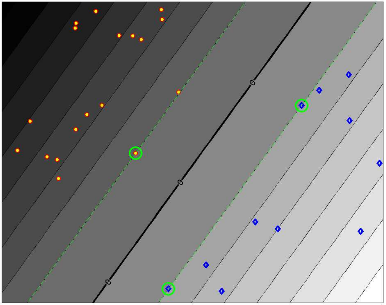


Data point influence

- $\alpha_i = 0$ this point is useless
- $\alpha_i \neq 0$ this point is said to be support

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{n} \alpha_i \, y_i(\mathbf{x}^\top \mathbf{x}_i) + b$$

Solving the dual (1)



Data point influence

- $\alpha_i = 0$ this point is useless
- $\alpha_i \neq 0$ this point is said to be support

$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \sum_{i=1}^{3} \alpha_i \, y_i(\mathbf{x}^\top \mathbf{x}_i) + b$$

Decison border only depends on 3 points ($d+1$)

Solving the dual (2)

Assume we know these 3 data points

$$\begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \\ \text{and} & 0 \leq \alpha_i \qquad\qquad i = 1, n \end{cases} \quad \Longrightarrow \quad \begin{cases} \min\limits_{\alpha \in \mathbb{R}^3} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{with} & \mathbf{y}^\top\alpha = 0 \end{cases}$$

$$L(\alpha, b) = \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha + b\, y^\top\alpha$$

solve the following linear system

$$\begin{cases} G\alpha + \quad b\,\mathbf{y} & = \mathbf{e} \\ \mathbf{y}^\top\alpha & = 0 \end{cases}$$

```
U = chol(G);  % upper
a = U\ (U'\e);
c = U\ (U'\y);
b = (y'*a)\(y'*c)
alpha = U\ (U'\(e - b*y));
```

Conclusion: variables or data point?

- seeking for a universal learning algorithm
  - no model for $\mathbb{P}(\mathbf{x}, y)$
- the linear case: data is separable
  - the non separable case
- double objective: minimizing the error together with the regularity of the solution
  - multi objective optimisation
- dualiy : variable – example
  - use the primal when $d < n$ (in the liner case) or when matrix G is hard to compute
  - otherwise use the dual
- universality = nonlinearity
  - kernels