Machine Learning Algorithms
Support Vector Machines and duality

Caio Corro

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,
91400, Orsay, France

Lagrangian relaxation and duality

Lagrangian relaxation 1/2

Let $U \subseteq \mathbb{R}^n$ be a convex set and $f : U \to \mathbb{R}$ be a function.
Primal mathematical program:

$$
\begin{aligned}
(P) \quad \min_{\boldsymbol{u} \in U} \quad & f(\boldsymbol{u}) \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{u} \leq \boldsymbol{b}
\end{aligned}
$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ defines a set of m linear inequalities.

Lagrangian relaxation 1/2

Let $U \subseteq \mathbb{R}^n$ be a convex set and $f : U \to \mathbb{R}$ be a function.
Primal mathematical program:

$$(P) \quad \min_{\boldsymbol{u} \in U} \quad f(\boldsymbol{u})$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{u} \leq \boldsymbol{b}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ defines a set of m linear inequalities.

Lagrangian

The Lagrangian of (P) is the function $L : U \times \mathbb{R}^m_+ \to \mathbb{R}$ defined as follows:

$$L(\boldsymbol{u}, \boldsymbol{\lambda}) = f(\boldsymbol{u}) + \boldsymbol{\lambda}^\top (\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b})$$

- $\boldsymbol{u} \in U$: primal variables
- $\boldsymbol{\lambda} \in \mathbb{R}^m_+$, i.e. $\boldsymbol{\lambda} \geq 0$: dual variables / Lagrangian multipliers

Lagrangian relaxation 2/2

Relaxation of the primal problem

The relaxed problem $L : \mathbb{R}_+^m \to \mathbb{R}$ is defined as:

$$L(\boldsymbol{\lambda}) = \min_{\boldsymbol{u} \in U} L(\boldsymbol{u}, \boldsymbol{\lambda})$$

Lagrangian relaxation 2/2

Relaxation of the primal problem

The relaxed problem $L : \mathbb{R}_+^m \to \mathbb{R}$ is defined as:

$$L(\boldsymbol{\lambda}) = \min_{\boldsymbol{u} \in U} L(\boldsymbol{u}, \boldsymbol{\lambda})$$

Weak Lagrangian duality (lower bound)

Let $\widehat{\boldsymbol{u}}$ be the optimal solution of the primal problem (P). Then:

$$\forall \boldsymbol{\lambda} \in \mathbb{R}_+^m : \quad f(\widehat{\boldsymbol{u}}) \geq L(\boldsymbol{\lambda})$$

Proof:

$$\begin{aligned} f(\widehat{\boldsymbol{u}}) &\geq f(\widehat{\boldsymbol{u}}) + \boldsymbol{\lambda}^\top (\boldsymbol{A}\widehat{\boldsymbol{u}} - b) \quad \text{(because } \widehat{\boldsymbol{u}} \text{ is satisfies the constraints)} \\ &\geq \min_{\boldsymbol{u} \in U} f(\boldsymbol{u}) + \boldsymbol{\lambda}^\top (\boldsymbol{A}\boldsymbol{u} - b) \\ &= L(\boldsymbol{\lambda}) \end{aligned}$$

Lagrangian dual problem

Definition

- $L(\boldsymbol{\lambda})$ is a lower bound to the primal problem, $\forall \boldsymbol{\lambda} \in \mathbb{R}^m_+ : \quad f(\widehat{\boldsymbol{u}}) \geq L(\boldsymbol{\lambda})$
- The dual problem search for the best lower bound

$$(D) \quad \max_{\boldsymbol{\lambda} \in \mathbb{R}^m_+} L(\boldsymbol{\lambda})$$

Lagrangian dual problem

Definition

- $L(\boldsymbol{\lambda})$ is a lower bound to the primal problem, $\forall \boldsymbol{\lambda} \in \mathbb{R}_+^m : \quad f(\widehat{\boldsymbol{u}}) \geq L(\boldsymbol{\lambda})$
- The dual problem search for the best lower bound

$$(D) \quad \max_{\boldsymbol{\lambda} \in \mathbb{R}_+^m} L(\boldsymbol{\lambda})$$

Concavity

The dual problem is concave (no matter if f is convex or not)

Strong Lagrangian duality

Let $\boldsymbol{\lambda} \in \mathbb{R}_+^m$ be a dual feasible solution and $\bar{\boldsymbol{u}} = \arg\max_{\boldsymbol{u} \in \mathcal{U}} L(\boldsymbol{u}, \boldsymbol{\lambda})$. If:

- $\boldsymbol{A}\bar{\boldsymbol{u}} \leq \boldsymbol{b}$ (primal feasibility condition)
- and $\boldsymbol{\lambda}^\top(\boldsymbol{A}\bar{\boldsymbol{u}} - \boldsymbol{b}) = 0$ (complementary slackness condition)

then $\bar{\boldsymbol{u}} = \widehat{\boldsymbol{u}}$ is a primal optimal solution.

Dual concavity proof 1/2

$L(\boldsymbol{\lambda})$ is concave if and only if:

1. the domain of $L(\boldsymbol{\lambda})$ is convex (trivial)
2. $\forall \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \boldsymbol{\lambda} \in \mathbb{R}_+^m, \epsilon \in [0,1] : L(\epsilon\boldsymbol{\lambda}^{(1)} + (1-\epsilon)\boldsymbol{\lambda}^{(2)}) \geq \epsilon L(\boldsymbol{\lambda}^{(1)}) + (1-\epsilon)L(\boldsymbol{\lambda}^{(2)})$

Dual concavity proof 1/2

$L(\boldsymbol{\lambda})$ is concave if and only if:

1. the domain of $L(\boldsymbol{\lambda})$ is convex (trivial)
2. $\forall \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \boldsymbol{\lambda} \in \mathbb{R}_+^m, \epsilon \in [0,1] : L(\epsilon \boldsymbol{\lambda}^{(1)} + (1-\epsilon)\boldsymbol{\lambda}^{(2)}) \geq \epsilon L(\boldsymbol{\lambda}^{(1)}) + (1-\epsilon) L(\boldsymbol{\lambda}^{(2)})$

Let $\boldsymbol{\lambda} = \epsilon \boldsymbol{\lambda}^{(1)} + (1-\epsilon)\boldsymbol{\lambda}^{(2)}$ and $\bar{\boldsymbol{u}} = \arg\min_{\boldsymbol{u} \in \mathcal{X}} L(\boldsymbol{u}, \boldsymbol{\lambda})$:

$$\left.\begin{array}{l} L(\bar{\boldsymbol{u}}, \boldsymbol{\lambda}^{(1)}) \geq L(\boldsymbol{\lambda}^{(1)}) \\ L(\bar{\boldsymbol{u}}, \boldsymbol{\lambda}^{(2)}) \geq L(\boldsymbol{\lambda}^{(2)}) \end{array}\right\} \Rightarrow \epsilon L(\bar{\boldsymbol{u}}, \boldsymbol{\lambda}^{(1)}) + (1-\epsilon) L(\bar{\boldsymbol{u}}, \boldsymbol{\lambda}^{(2)}) \geq \epsilon L(\boldsymbol{\lambda}^{(1)}) + (1-\epsilon) L(\boldsymbol{\lambda}^{(2)})$$

To understand the left-hand side, remember that:

$$L(\boldsymbol{\lambda}) = \min_{\boldsymbol{u} \in \mathcal{X}} L(\boldsymbol{u}, \boldsymbol{\lambda})$$

So:

- the right-hand side has the expected form
- the left-hand side is different, we need to fix this to finish the proof

Dual concavity proof 2/2

To simplify notations, we write $c(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}$.

The left-hand side of the inequality can be rewritten as:

$$\begin{aligned}
\epsilon L(\bar{\boldsymbol{u}}, \boldsymbol{\lambda}^{(1)}) + (1-\epsilon) L(\bar{\boldsymbol{u}}, \boldsymbol{\lambda}^{(2)}) &= \epsilon \left(f(\bar{\boldsymbol{u}}) + \boldsymbol{\lambda}^{(1)\top} c(\bar{\boldsymbol{u}}) \right) + (1-\epsilon) \left(f(\bar{\boldsymbol{u}}) + \boldsymbol{\lambda}^{(2)\top} c(\bar{\boldsymbol{u}}) \right) \\
&= \epsilon f(\bar{\boldsymbol{u}}) + \epsilon \boldsymbol{\lambda}^{(1)\top} c(\bar{\boldsymbol{u}}) + (1-\epsilon) f(\bar{\boldsymbol{u}}) + (1-\epsilon) \boldsymbol{\lambda}^{(2)\top} c(\bar{\boldsymbol{u}}) \\
&= f(\bar{\boldsymbol{u}}) + \left(\epsilon \boldsymbol{\lambda}^{(1)} + (1-\epsilon) \boldsymbol{\lambda}^{(2)} \right)^\top c(\bar{\boldsymbol{u}}) \\
&= f(\bar{\boldsymbol{u}}) + \boldsymbol{\lambda}^\top c(\bar{\boldsymbol{u}}) \\
&= L(\bar{\boldsymbol{u}}, \boldsymbol{\lambda}) \\
&= L(\boldsymbol{\lambda}) \\
&= L(\epsilon \boldsymbol{\lambda}^{(1)} + (1-\epsilon) \boldsymbol{\lambda}^{(2)})
\end{aligned}$$

Hence, we obtain the inequality:

$$L(\epsilon \boldsymbol{\lambda}^{(1)} + (1-\epsilon) \boldsymbol{\lambda}^{(2)}) \geq \epsilon L(\boldsymbol{\lambda}^{(1)}) + (1-\epsilon) L(\boldsymbol{\lambda}^{(2)})$$

which proves that the Lagrangian dual objective is concave.

Strong Lagrangian duality proof

Let $\widehat{\boldsymbol{u}}$ be a primal optimal solution.
By weak Lagrangian duality, we know that:

$$\begin{aligned} f(\widehat{\boldsymbol{u}}) &\geq L(\boldsymbol{\lambda}) \\ &= L(\bar{\boldsymbol{u}}, \boldsymbol{\lambda}) \\ &= f(\bar{\boldsymbol{u}}) + \boldsymbol{\lambda}^\top (\boldsymbol{A}\bar{\boldsymbol{u}} - \boldsymbol{b}) \end{aligned}$$

From the prerequisites (complementary slackness) the second term in null:

$$f(\widehat{\boldsymbol{u}}) \geq f(\bar{\boldsymbol{u}})$$

Moreover: $f(\widehat{\boldsymbol{u}}) \leq f(\widehat{\boldsymbol{u}})$ because $\widehat{\boldsymbol{u}}$ is primal feasible,
Therefore $f(\bar{\boldsymbol{u}}) = f(\widehat{\boldsymbol{u}})$.

Relaxing equality constraints

$$
\begin{aligned}
\min_{\boldsymbol{u}} \quad & f(\boldsymbol{u}) \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{u} = \boldsymbol{b}
\end{aligned}
\qquad \Leftrightarrow \qquad
\begin{aligned}
\min_{\boldsymbol{u}} \quad & f(\boldsymbol{u}) \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{u} \leq \boldsymbol{b} \\
& -\boldsymbol{A}\boldsymbol{u} \leq -\boldsymbol{b}
\end{aligned}
$$

Relaxing equality constraints

$$\min_{\boldsymbol{u}} \quad f(\boldsymbol{u}) \qquad\qquad \Leftrightarrow \qquad \min_{\boldsymbol{u}} \quad f(\boldsymbol{u})$$

$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{u} = \boldsymbol{b} \qquad\qquad\qquad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{u} \leq \boldsymbol{b}$$

$$-\boldsymbol{A}\boldsymbol{u} \leq -\boldsymbol{b}$$

Lagrangian:
$$\begin{aligned} L(\boldsymbol{u}, \boldsymbol{\lambda} \geq 0, \boldsymbol{\lambda}' \geq 0) &= f(\boldsymbol{u}) + \boldsymbol{\lambda}^\top(\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}) + \boldsymbol{\lambda}'^\top(-\boldsymbol{A}\boldsymbol{u} + \boldsymbol{b}) \\ &= f(\boldsymbol{u}) + \boldsymbol{\lambda}^\top(\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}) - \boldsymbol{\lambda}'^\top(\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}) \\ &= f(\boldsymbol{u}) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}')^\top(\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}) \end{aligned}$$

Let $\boldsymbol{\lambda}'' = \boldsymbol{\lambda} - \boldsymbol{\lambda}'$:

$$L(\boldsymbol{u}, \boldsymbol{\lambda}'') = f(\boldsymbol{u}) + \boldsymbol{\lambda}''^\top(\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b})$$

In this formulation, the dual variables $\boldsymbol{\lambda}'' \in \mathbb{R}^m$ is **unconstrained**.

Benefits of relaxing equalities

Unconstrained optimization problem + Simpler strong duality condition

Optimality conditions
Constrained optimization problems

KKT conditions for a minimization problems 1/3

Primal problem (MINIMIZATION)

$$\begin{aligned}
\min_{\boldsymbol{u} \in R^k} \quad & f(\boldsymbol{u}) \\
\text{s.t.} \quad & g^{(i)}(\boldsymbol{u}) \leq 0 & \forall 1 \leq i \leq m \\
& h^{(i)}(\boldsymbol{u}) = 0 & \forall 1 \leq i \leq n
\end{aligned}$$

- $g^{(i)} \leq 0$: a set of m inequality constraints
- $h^{(i)}(\boldsymbol{u}) = 0$: a set of n equality constraints

Lagrangian

$$L(\boldsymbol{u}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{u}) + \sum_j \mu_j g^{(j)}(\boldsymbol{u}) + \sum_j \lambda_j h^{(j)}(\boldsymbol{u})$$

- $\boldsymbol{\mu} \in \mathbb{R}_+^m$: dual variables associated with primal inequalities
- $\boldsymbol{\lambda} \in \mathbb{R}^n$: dual variables associated with primal equalities

KKT conditions for a minimization problem 2/3

Lagrangian

$$L(\boldsymbol{u}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{\mu}) + \sum_j \mu_j g^{(j)}(\boldsymbol{u}) + \sum_j \lambda_j h^{(j)}(\boldsymbol{u})$$

Necessary optimal condition

Any optimal primal/dual triplet $(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\lambda}})$ satisfies the following conditions:

$$\begin{array}{lll}
\text{(stationarity)} & \forall i: & \partial_{\widehat{\boldsymbol{u}}} \left(f(\widehat{\boldsymbol{u}}) + \sum_j \widehat{\mu}_j g^{(j)}(\widehat{\boldsymbol{u}}) + \sum_j \widehat{\lambda}_j h^{(j)}(\widehat{\boldsymbol{u}}) \right) \ni \mathbf{0} \\
\text{(primal feasibility)} & \forall i: & g^{(i)}(\widehat{\boldsymbol{u}}) \leq 0 \\
 & \forall i: & h^{(i)}(\widehat{\boldsymbol{u}}) = 0 \\
\text{(dual feasibility)} & \forall i: & \widehat{\mu}_i \geq 0 \\
\text{(complementary slackness)} & & \sum_i \widehat{\mu}_i g^{(i)}(\widehat{\boldsymbol{u}}) = 0
\end{array}$$

KKT conditions for a minimization problem 3/3

Warning

KKT conditions are necessary constraints:
there may exists points $\boldsymbol{u}, \boldsymbol{\mu}, \boldsymbol{\lambda}$ that satisfies these constraints that are not optimal solution

Necessary and sufficient conditions

If:

- f is a concave function,
- all $g^{(i)}$ are convex functions,
- and all $h^{(i)}$ are affine functions,

then the KKT are sufficient conditions.
$=>$ we can solve this set of equations to find the optimal $\widehat{\boldsymbol{u}}$, i.e. the optimization problem (may) have a closed form solution

KKT conditions for linear constraints 1/2

Primal problem

$$\begin{aligned} \min_{\boldsymbol{u} \in R^k} \quad & f(\boldsymbol{u}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{u} \leq \boldsymbol{b} \\ & \boldsymbol{C}\boldsymbol{u} = \boldsymbol{d} \end{aligned}$$

Lagrangian

$$L(\boldsymbol{u}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{u}) + \langle \boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{C}\boldsymbol{u} - \boldsymbol{d} \rangle$$

- $\boldsymbol{\mu} \in \mathbb{R}_+^m$: dual variables associated with primal inequalities
- $\boldsymbol{\lambda} \in \mathbb{R}^n$: dual variables associated with primal equalities

KKT conditions for linear constraints 2/2

Lagrangian

$$L(\boldsymbol{u}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{u}) + \langle \boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{C}\boldsymbol{u} - \boldsymbol{d} \rangle$$

Necessary and sufficient optimality condition

An optimal primal/dual triplet $(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\lambda}})$ satisfies the following constraints:

(stationarity)	$\partial_{\widehat{\boldsymbol{u}}} \Big(f(\widehat{\boldsymbol{u}}) + \langle \widehat{\boldsymbol{\mu}}, \boldsymbol{A}\widehat{\boldsymbol{u}} - \boldsymbol{b} \rangle + \langle \widehat{\boldsymbol{\lambda}}, \boldsymbol{C}\widehat{\boldsymbol{u}} - \boldsymbol{d} \rangle \Big) \ni \boldsymbol{0}$
(primal feasibility)	$\boldsymbol{A}\widehat{\boldsymbol{u}} \leq \boldsymbol{b}$
	$\boldsymbol{C}\widehat{\boldsymbol{u}} = \boldsymbol{d}$
(dual feasibility)	$\widehat{\boldsymbol{\mu}} \geq 0$
(complementary slackness)	$\langle \widehat{\boldsymbol{\mu}}, \boldsymbol{A}\widehat{\boldsymbol{u}} - \boldsymbol{b} \rangle = 0$

KKT conditions for a maximization problem 1/3

Primal problem (MAXIMIZATION)

$$\begin{aligned} \max_{\boldsymbol{u}\in R^k} \quad & f(\boldsymbol{u}) & \\ \text{s.t.} \quad & g^{(i)}(\boldsymbol{u}) \geq 0 & \forall 1 \leq i \leq m \\ & h^{(i)}(\boldsymbol{u}) = 0 & \forall 1 \leq i \leq n \end{aligned}$$

- $g^{(i)} \geq 0$: a set of m inequality constraints
- $h^{(i)}(\boldsymbol{u}) = 0$: a set of n equality constraints

Lagrangian

$$L(\boldsymbol{u}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{u}) + \sum_j \mu_j g^{(j)}(\boldsymbol{u}) + \lambda_j \sum_j h^{(j)}(\boldsymbol{u})$$

- $\boldsymbol{\mu} \in \mathbb{R}^m_+$: dual variables associated with primal inequalities
- $\boldsymbol{\lambda} \in \mathbb{R}^n$: dual variables associated with primal equalities

Fenchel duality

Let $f : \mathbb{R}^m \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{U} \in \mathbb{R}^{m \times n}$ and assume the following problem:

$$\min_{\boldsymbol{v} \in \mathbb{R}^n} f(\boldsymbol{U}\boldsymbol{v}) + h(\boldsymbol{v})$$

Let $f : \mathbb{R}^m \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{U} \in \mathbb{R}^{m \times n}$ and assume the following problem:

$$\min_{\boldsymbol{v} \in \mathbb{R}^n} f(\boldsymbol{U}\boldsymbol{v}) + h(\boldsymbol{v})$$

We can rewrite the problem by introducing a term $\boldsymbol{t} \in \mathbb{R}^m$:

$$= \min_{\boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{t} \in \mathbb{R}^m} f(\boldsymbol{t}) + h(\boldsymbol{v}) \quad \text{s.t.} \quad \boldsymbol{t} = \boldsymbol{U}\boldsymbol{v}$$

Let $f : \mathbb{R}^m \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{U} \in \mathbb{R}^{m \times n}$ and assume the following problem:

$$\min_{\boldsymbol{v} \in \mathbb{R}^n} f(\boldsymbol{U}\boldsymbol{v}) + h(\boldsymbol{v})$$

We can rewrite the problem by introducing a term $\boldsymbol{t} \in \mathbb{R}^m$:

$$= \min_{\boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{t} \in \mathbb{R}^m} f(\boldsymbol{t}) + h(\boldsymbol{v}) \quad \text{s.t.} \quad \boldsymbol{t} = \boldsymbol{U}\boldsymbol{v}$$

We relax the constraint using dual vars $\boldsymbol{\lambda} \in \mathbb{R}^m$ and build the Lagrangian dual:

$$\geq \max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \min_{\boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{t} \in \mathbb{R}^m} f(\boldsymbol{t}) + h(\boldsymbol{v}) + \boldsymbol{\lambda}^\top (\boldsymbol{U}\boldsymbol{v} - \boldsymbol{t})$$

Let $f : \mathbb{R}^m \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{U} \in \mathbb{R}^{m\times n}$ and assume the following problem:

$$\min_{\boldsymbol{v}\in\mathbb{R}^n} f(\boldsymbol{U}\boldsymbol{v}) + h(\boldsymbol{v})$$

We can rewrite the problem by introducing a term $\boldsymbol{t} \in \mathbb{R}^m$:

$$= \min_{\boldsymbol{v}\in\mathbb{R}^n, \boldsymbol{t}\in\mathbb{R}^m} f(\boldsymbol{t}) + h(\boldsymbol{v}) \quad \text{s.t.} \quad \boldsymbol{t} = \boldsymbol{U}\boldsymbol{v}$$

We relax the constraint using dual vars $\boldsymbol{\lambda} \in \mathbb{R}^m$ and build the Lagrangian dual:

$$\geq \max_{\boldsymbol{\lambda}\in\mathbb{R}^m} \min_{\boldsymbol{v}\in\mathbb{R}^n, \boldsymbol{t}\in\mathbb{R}^m} f(\boldsymbol{t}) + h(\boldsymbol{v}) + \boldsymbol{\lambda}^\top(\boldsymbol{U}\boldsymbol{v} - \boldsymbol{t})$$

We re-arrange terms as follows:

$$= \max_{\boldsymbol{\lambda}\in\mathbb{R}^m} \left(\min_{\boldsymbol{t}\in\mathbb{R}^m} f(\boldsymbol{t}) - \boldsymbol{\lambda}^\top \boldsymbol{t}\right) + \left(\min_{\boldsymbol{v}\in\mathbb{R}^n} h(\boldsymbol{v}) + \boldsymbol{\lambda}^\top \boldsymbol{U}\boldsymbol{v}\right)$$

$$= \max_{\boldsymbol{\lambda}\in\mathbb{R}^m} -\left(\max_{\boldsymbol{t}\in\mathbb{R}^m} \boldsymbol{\lambda}^\top \boldsymbol{t} - f(\boldsymbol{t})\right) - \left(\max_{\boldsymbol{v}\in\mathbb{R}^n} -\boldsymbol{\lambda}^\top \boldsymbol{U}\boldsymbol{v} - h(\boldsymbol{v})\right)$$

Fenchel duality

Let $f : \mathbb{R}^m \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{U} \in \mathbb{R}^{m \times n}$ and assume the following problem:

$$\min_{\boldsymbol{v} \in \mathbb{R}^n} f(\boldsymbol{U}\boldsymbol{v}) + h(\boldsymbol{v})$$

Then the Fenchel dual problem is defined as follows:

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} -f^*(\boldsymbol{\lambda}) - h^*(-\boldsymbol{\lambda}^\top \boldsymbol{U})$$

- Under certain conditions, the two problem have the same solution
- Using stationarity condition from the Lagrangian dual, we can retrieve primal-dual relationship (example later)

Support vector machines for binary classification

Binary classification

Task

Given a vector of feature values $\boldsymbol{x} \in \mathbb{R}^d$ we want to predict $y \in \{-1, 1\}$, i.e. either class -1 or 1.

$$y = \begin{cases} 1 & \text{if } \langle \boldsymbol{a}, \boldsymbol{x} \rangle \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

=> predict the class which has the same sign as $\langle \boldsymbol{a}, \boldsymbol{x} \rangle$ (the choice of setting $\langle \boldsymbol{a}, \boldsymbol{x} \rangle = 0$ to class 1 is arbitrary).

In our framework

$$s_{\boldsymbol{a}}(\boldsymbol{u}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle$$

$$\widehat{y}(w) = \begin{cases} 1 & \text{if } w \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Support Vector Machine (SVM)

SVM training problem

Given a training set D of n examples, $(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})$
$=>$ find an hyperplane that separates the two classes

$$\begin{aligned} \min_{\boldsymbol{a}} \quad & \alpha r(\boldsymbol{a}) \\ \text{s.t.} \quad & \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)} \geq m \qquad \forall (\boldsymbol{x}^{(i)}, y^{(i)}) \in D \end{aligned}$$

- m: minimum required margin, usually $m = 1$
- $r(\boldsymbol{a}) = 0 =>$ any hyperplane (no objective function!)
- $r(\boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{a}\|_2^2 =>$ hyperplane that maximizes the margin [Boser et al., 1992, Section 2.1]

SVM reformulation 1/3

Downside the previous SVM formulation

- constrained optimization is difficult
- what about non separable data?

SVM reformulation 1/3

Downside the previous SVM formulation

- constrained optimization is difficult
- what about non separable data?

SVM with slack variables

$$
\begin{aligned}
\min_{\boldsymbol{a},\epsilon} \quad & \alpha r(\boldsymbol{a}) + \sum_{i=1}^{n} \epsilon_i \\
\text{s.t.} \quad & \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)} \geq m - \epsilon_i \qquad (\boldsymbol{x}^{(i)}, y^{(i)}) \in D \\
& \epsilon \geq 0
\end{aligned}
$$

- $\epsilon \in \mathbb{R}^n$: the vector of slack variables, one slack variable per datapoint
- one slack variable per datapoint
- optimize both over $\boldsymbol{a}$ and ϵ.

SVM reformulation 2/3

SVM with slack variables

$$
\begin{aligned}
\min_{\boldsymbol{a},\epsilon} \quad & \alpha r(\boldsymbol{a}) + \sum_{i=1}^{n} \epsilon_i \\
\text{s.t.} \quad & \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)} \geq m - \epsilon_i \qquad (\boldsymbol{x}^{(i)}, y^{(i)}) \in D \\
& \epsilon \geq 0
\end{aligned}
$$

Constraints reformulation

$$
\begin{aligned}
\langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)} &\geq m - \epsilon_i & \forall (\boldsymbol{x}^{(i)}, y^{(i)}) \in D \\
\langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)} - m &\geq -\epsilon_i & \forall (\boldsymbol{x}^{(i)}, y^{(i)}) \in D \\
m - \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)} &\leq \epsilon_i & \forall (\boldsymbol{x}^{(i)}, y^{(i)}) \in D \\
\max(0, m - \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)}) &= \epsilon_i & \forall (\boldsymbol{x}^{(i)}, y^{(i)}) \in D
\end{aligned}
$$

SVM reformulation 3/3

SVM with slack variables

$$
\begin{aligned}
\min_{\boldsymbol{a},\epsilon} \quad & \alpha r(\boldsymbol{a}) + \sum_{i=1}^{n} \epsilon_i \\
\text{s.t.} \quad & \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)} \geq m - \epsilon_i \quad \Leftrightarrow \quad \max(0, m - \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)}) = \epsilon_i \qquad \forall (\boldsymbol{x}^{(i)}, y^{(i)}) \in D \\
& \epsilon \geq 0
\end{aligned}
$$

Unconstrained SVM training problem

$$
\begin{aligned}
& \min_{\boldsymbol{a}} \quad \sum_{i=1}^{n} \max(0, m - \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)}) + \alpha r(\boldsymbol{a}) \\
= & \min_{\boldsymbol{a}} \quad \sum_{i=1}^{n} \ell_m(y^{(i)}, \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle) + \alpha r(\boldsymbol{a})
\end{aligned}
$$

Remark

- if $m > 0$: hinge loss (usually set $m = 1$)
- if $m = 0$: perceptron loss

where $\ell_m(y, w) = \max(0, m - yw)$ is the binary hinge loss function, parameterized by the margin parameter $m \geq 0$

Support vector machines for multiclass classification

Multiclass classification

Task

- Predicting a 1-in-k class given a set of feature values $\boldsymbol{x}$
- Usually $k > 2$, but works for $k = 2$

In our framework

Let d be the number of features and k the number of classes.

- Scoring function: $s(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ where $\boldsymbol{A} \in \mathbb{R}^{k \times d}, \boldsymbol{b} \in \mathbb{R}^{d}$
- Non-probabilistic prediction function: $\widehat{y}(\boldsymbol{w}) = \arg\max_{\boldsymbol{y} \in E(k)} \langle \boldsymbol{w}, \boldsymbol{y} \rangle$
- Probabilistic prediction function: $\widehat{y}(\boldsymbol{w}) = \mathrm{softmax}(\boldsymbol{w})$

Support Vector Machine (SVM)

Given a training set D of n datapoints $(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}), \ldots, (\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$ the SVM training prolem is defined as follows:

$$\begin{aligned} \arg\min_{\boldsymbol{A}} \quad & \alpha r(\boldsymbol{A}) \\ \text{s.t.} \quad & \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle + m \leq \langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \qquad \forall (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in D, \boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\} \end{aligned}$$

- $k - 1$ constraints per training point with this formulation
- if highest scoring class that is not the gold class satisfies the constraint, all other non-gold classes will also satisfy it

Support Vector Machine (SVM)

Given a training set D of n datapoints $(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}), \ldots, (\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$ the SVM training prolem is defined as follows:

$$\begin{aligned} \arg\min_{\boldsymbol{A}} \quad & \alpha r(\boldsymbol{A}) \\ \text{s.t.} \quad & \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle + m \leq \langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \qquad \forall (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in D, \boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\} \end{aligned}$$

- $k - 1$ constraints per training point with this formulation
- if highest scoring class that is not the gold class satisfies the constraint, all other non-gold classes will also satisfy it

Alternative constraint set:

$$m + \max_{\boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\}} \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \leq \langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \qquad \forall (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in D$$

Slack variable reformulation

Constraints:

$$m + \max_{\boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\}} \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \leq \langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \qquad \forall (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in D$$

Add slack variables $\epsilon \in \mathbb{R}_+^n$ in case the problem is not separable:

$$m + \max_{\boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\}} \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \leq \langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle + \epsilon_i \qquad \forall (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in$$

$$-\langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle + m + \max_{\boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\}} \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \leq \epsilon_i \qquad \forall (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in D$$

$$\max \left(0, -\langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle + m + \max_{\boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\}} \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \right) = \epsilon_i \qquad \forall (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in D$$

The two SVM variants

Constrained optimization problem

$$
\begin{aligned}
\underset{\boldsymbol{A}}{\arg\min} \quad & \alpha r(\boldsymbol{A}) + \sum_{i=1}^{n} \epsilon_i \\
\text{s.t.} \quad & \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle + m \leq \langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle + \epsilon \qquad \forall (\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in D, \boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\} \\
& \epsilon \geq 0
\end{aligned}
$$

Unconstrained optimization problem

$$
\underset{\boldsymbol{A}}{\arg\min} \quad \sum_{i=1}^{n} \max \left(0, -\langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle + m + \max_{\boldsymbol{y}' \in \mathcal{Y}(k) \setminus \{\boldsymbol{y}^{(i)}\}} \langle \boldsymbol{y}', \boldsymbol{A}\boldsymbol{x}^{(i)} \rangle \right) + \alpha r(\boldsymbol{A})
$$

Dual SVM training problem

Binary SVM with slack variables

SVM with slack variables

$$\begin{aligned}
\min_{\boldsymbol{a},\epsilon} \quad & \alpha r(\boldsymbol{a}) + \sum_{i=1}^{n} \epsilon_i \\
\text{s.t.} \quad & \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)} \geq m - \epsilon_i \quad \Leftrightarrow \quad \max(0, m - \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)}) = \epsilon_i \qquad \forall (\boldsymbol{x}^{(i)}, y^{(i)}) \in D \\
& \epsilon \geq 0
\end{aligned}$$

Unconstrained SVM training problem

$$\begin{aligned}
\min_{\boldsymbol{a}} \quad & \sum_{i=1}^{n} \max(0, m - \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle y^{(i)}) + \alpha r(\boldsymbol{a}) \\
= \min_{\boldsymbol{a}} \quad & \sum_{i=1}^{n} \ell_m(y^{(i)}, \langle \boldsymbol{a}, \boldsymbol{x}^{(i)} \rangle) + \alpha r(\boldsymbol{a})
\end{aligned}$$

Remark

- if $m > 0$: hinge loss (usually set $m = 1$)
- if $m = 0$: perceptron loss

where $l_m(y, w) = \max(0, m - yw)$ is the binary hinge loss function, parameterized by the margin parameter $m \geq 0$

Motivations

SVM training problem

$$\min_{\boldsymbol{a}} \sum_{(\boldsymbol{x},\boldsymbol{y})\in D} \max(0, m - \langle \boldsymbol{x}, \boldsymbol{x} \rangle y) + \frac{1}{2}\|\boldsymbol{a}\|_2^2$$

The hinge loss is not differentiable everywhere:

$$\ell_m(y, w) = \max(0, m - wy)$$

SVM dual formulation benefits

- Differentiable! => no need for a subgradient
- Can be trained with a hyper-parameter free algorithm! => no step-size

Notation change

Notation change

- $\boldsymbol{X} \in \mathbb{R}^{n \times d}$: matrix where each row consists of a training point, i.e. $X_{i,j} = x_j^{(i)}$
- $\boldsymbol{Y} \in R^{n \times n}$: diagonal matrix containing labels, i.e. $Y_{i,i} = y^{(i)}$ and $\forall i \neq j : Y_{i,j} = 0$.
- $\ell(v_i) = \max(0, 1 - v_i)$

$$\min_{\boldsymbol{a},\boldsymbol{v}} \quad \sum_{i=1}^{n} \ell([\boldsymbol{YXa}]_i) + \frac{1}{2}\|\boldsymbol{a}\|_2^2$$

Fenchel conjugate of the Hinge loss 1/2

Conjugate of separable functions [Beck, Theorem 4.12]

Let $h : \mathbb{R}^{\boldsymbol{n}} \to \mathbb{R} \cup \{\infty\}$ and $f_i : \mathbb{R} \to \mathbb{R} \cup \{\infty\}, i \in \{1...n\}$ be functions defined as:

$$h(\boldsymbol{v}) = \sum_i f_i(v_i)$$

Then, the conjugate of h is defined as:

$$h^*(\boldsymbol{u}) = \sum_i f_i^*(u_i)$$

Hinge loss

The Hinge loss term in the SVM object is a separable function:

$$\sum_{i=1}^{n} max(0, 1 - [\boldsymbol{YXa}]_i)$$

Fenchel conjugate of the Hinge loss 2/2

Hinge loss

The Hinge loss term in the SVM object is a separable function:

$$\sum_{i=1}^{n} max(0, 1 - [\boldsymbol{YXa}]_i)$$

Hinge loss conjugate [Beck, Section 4.4.3]

Let $f(w) = \max(0, 1 - w)$, then $f^*(u) = u + \delta_{[-1,0]}(u)$.

Note: The indicator function act as constraint on the domain of the dual variables in the objective

Fenchel dual of the SVM training problem 1/2

Fenchel dual problem

Let $f : \mathbb{R}^m \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{U} \in \mathbb{R}^{m \times n}$ and assume the following problem:

$$\min_{\boldsymbol{v} \in \mathbb{R}^n} f(\boldsymbol{U}\boldsymbol{v}) + h(\boldsymbol{v})$$
$$\geq \max_{\boldsymbol{\lambda} \in \mathbb{R}^m} -f^*(\boldsymbol{\lambda}) - h^*(-\boldsymbol{U}^\top \boldsymbol{\lambda})$$

SVM case

- $\boldsymbol{U} = \boldsymbol{Y}\boldsymbol{X}$
- $\boldsymbol{v} = \boldsymbol{a}$

If d is the input dimension and n the number of datapoints:

$$\min_{\boldsymbol{a} \in \mathbb{R}^d} f(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{a}) + h(\boldsymbol{v})$$
$$\geq \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} -f^*(\boldsymbol{\lambda}) - h^*(-(\boldsymbol{Y}\boldsymbol{X})^\top \boldsymbol{\lambda})$$

Fenchel dual of the SVM training problem 2/2

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} -f^*(\boldsymbol{\lambda}) - h^*(-(\boldsymbol{YX})^\top \boldsymbol{\lambda})$$

- $h^*(-(\boldsymbol{YX})^\top \boldsymbol{\lambda})$: this term is trivial because h is quadratic regularization!
- $f^*(\boldsymbol{\lambda})$: Fenchel conjugate of the Hinge loss

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & -\sum_{i=1}^n \boldsymbol{\lambda}_i - \frac{1}{2} \boldsymbol{\lambda}^\top \boldsymbol{YXX}^\top \boldsymbol{Y\lambda} \\ \text{s.t.} \quad & -1 \leq \boldsymbol{\lambda}_i \leq 0 \quad \forall 1 \leq i \leq n \end{aligned}$$

Recovering optimal primal variables from optimal dual variables

From the KKT's stationarity conditions, for optimal primal and dual variables we have:

$$
\begin{aligned}
\nabla_{\boldsymbol{a}} L(\boldsymbol{a}, \boldsymbol{v}, \boldsymbol{\lambda}) &= 0 \\
\nabla_{\boldsymbol{a}} \left(\sum_{i=1}^{n} \max(0, 1 - v_i) + \frac{1}{2} \|\boldsymbol{a}\|_2^2 + \boldsymbol{\lambda}^\top (\boldsymbol{Y}\boldsymbol{X}\boldsymbol{a} - \boldsymbol{v}) \right) &= 0 \\
\boldsymbol{a} + \left(\boldsymbol{\lambda}^\top \boldsymbol{Y}\boldsymbol{X} \right)^\top &= 0 \\
\boldsymbol{a} &= -\boldsymbol{X}^\top \boldsymbol{Y} \boldsymbol{\lambda}
\end{aligned}
$$