Support vector machines

Guillaume Obozinski

Ecole des Ponts - ParisTech

SOCN course 2014

Outline

Outline

1. Constrained optimization, Lagrangian duality and KKT

2. Support vector machines

Constrained optimization, Lagrangian duality and KKT

Review: Constrained optimization

Optimization problem in canonical form

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & h_i(\mathbf{x}) = 0, \quad i \in [\![1, n]\!] \\ & g_j(\mathbf{x}) \leq 0, \quad j \in [\![1, m]\!] \end{aligned}$$

with

- $\mathcal{X} \subset \mathbb{R}^p$.
- f, g_j functions,
- h_i affine functions.

Review: Constrained optimization

Optimization problem in canonical form

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{X}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & h_i(\mathbf{x}) = 0, \quad i \in [\![1, n]\!] \\ & g_j(\mathbf{x}) \leq 0, \quad j \in [\![1, m]\!] \end{aligned}$$

with

- $\mathcal{X} \subset \mathbb{R}^p$.
- f, g_j functions,
- h_i affine functions.

The problem is convex if f, g_j and $\mathcal{X}$ are convex (w.l.o.g $\mathring{\mathcal{X}} \neq \varnothing$).

Review: Constrained optimization

Optimization problem in canonical form

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & h_i(\mathbf{x}) = 0, \quad i \in [\![1, n]\!] \\ & g_j(\mathbf{x}) \leq 0, \quad j \in [\![1, m]\!] \end{aligned}$$

with

- $\mathcal{X} \subset \mathbb{R}^p$.
- f, g_j functions,
- h_i affine functions.

The problem is convex if f, g_j and $\mathcal{X}$ are convex (w.l.o.g $\mathring{\mathcal{X}} \neq \varnothing$).

Lagrangian

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i \, h_i(\mathbf{x}) + \sum_{j=1}^{m} \mu_j \, g_j(\mathbf{x})$$

Lagrangian duality

Lagrangian

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i \, h_i(\mathbf{x}) + \sum_{j=1}^{m} \mu_j \, g_j(\mathbf{x})$$

Lagrangian duality

Lagrangian

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i \, h_i(\mathbf{x}) + \sum_{j=1}^{m} \mu_j \, g_j(\mathbf{x})$$

Primal vs Dual problem

$$p^* = \min_{\mathbf{x} \in \mathcal{X}} \quad \max_{\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m_+} \quad \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad \text{(P)}$$

$$d^* = \max_{\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m_+} \quad \min_{\mathbf{x} \in \mathcal{X}} \quad \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad \text{(D)}$$

Maxmin inequalities

$$\max_y \min_x f(x,y) \leq \min_x \max_y f(x,y)$$

Maxmin inequalities

$$f(x, y) \leq \qquad \max_y f(x, y)$$

Maxmin inequalities

$$\min_x f(x, y) \leq \min_x \max_y f(x, y)$$

Maxmin inequalities

$$\max_y \min_x f(x,y) \leq \min_x \max_y f(x,y)$$

Maxmin inequalities

$$\max_y \min_x f(x,y) \leq \min_x \max_y f(x,y)$$

Weak duality

In general, we have $d^* \leq p^*$. This is called weak duality.

Maxmin inequalities

$$\max_{y} \min_{x} f(x, y) \leq \min_{x} \max_{y} f(x, y)$$

Weak duality

In general, we have $d^* \leq p^*$. This is called weak duality.

Strong duality

In some cases, we have strong duality:

- $d^* = p^*$
- Solutions to (P) and (D) are the same

Slater's qualification condition

Slater's qualification condition is a condition on the constraints that guarantees that strong duality holds.

Consider an optimization problem in canonical form.

Slater's qualification condition

Slater's qualification condition is a condition on the constraints that guarantees that strong duality holds.

Consider an optimization problem in canonical form.

Definition: Slater's condition (strong form)

There exists $\mathbf{x} \in \mathring{\mathcal{X}}$ such that $h(\mathbf{x}) = 0$ and $g(\mathbf{x}) < 0$ entrywise.

Slater's qualification condition

Slater's qualification condition is a condition on the constraints that guarantees that strong duality holds.

Consider an optimization problem in canonical form.

Definition: Slater's condition (strong form)

There exists $\mathbf{x} \in \mathring{\mathcal{X}}$ such that $h(\mathbf{x}) = 0$ and $g(\mathbf{x}) < 0$ entrywise.

Definition: Slater's condition (weak form)

There exists $\mathbf{x} \in \mathring{\mathcal{X}}$ such that $h(\mathbf{x}) = 0$ and $g(\mathbf{x}) \leq 0$ entrywise, but with $g_i(\mathbf{x}) < 0$ if g_i is not affine.

Slater's qualification condition

Slater's qualification condition is a condition on the constraints that guarantees that strong duality holds.

Consider an optimization problem in canonical form.

Definition: Slater's condition (strong form)

There exists $\mathbf{x} \in \mathring{\mathcal{X}}$ such that $h(\mathbf{x}) = 0$ and $g(\mathbf{x}) < 0$ entrywise.

Definition: Slater's condition (weak form)

There exists $\mathbf{x} \in \mathring{\mathcal{X}}$ such that $h(\mathbf{x}) = 0$ and $g(\mathbf{x}) \leq 0$ entrywise, but with $g_i(\mathbf{x}) < 0$ if g_i is not affine.

Slater's conditions requires that there exists a feasible point which is strictly feasible for all non-affine constraints.

Karush-Kuhn-Tucker conditions

Theorem

For a convex problem defined by differentiable functions f, h_i, g_j, x is an optimal solution if and only if there exists $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ such that the KKT conditions are satisfied.

KKT conditions

$$\nabla f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i \nabla h_i(\mathbf{x}) + \sum_{j=1}^{m} \mu_j \nabla g_j(\mathbf{x}) = 0 \quad \text{(Lagrangian stationarity)}$$

$$h(\mathbf{x}) = 0, \quad g(\mathbf{x}) \leq 0 \quad \text{(primal feasibility)}$$

$$\mu_j \geq 0 \quad \text{(dual feasibility)}$$

$$\forall j \in [\![1, m]\!], \quad \mu_j \, g_j(\mathbf{x}) = 0 \quad \text{(complementary slackness)}$$

Outline

Support vector machines

Hard margin SVM

- Binary classification problem with $y_i \in \{-1, 1\}$.

Hard margin SVM

- Binary classification problem with $y_i \in \{-1, 1\}$.
- Margin $\frac{1}{\|w\|}$

Hard margin SVM

- Binary classification problem with $y_i \in \{-1, 1\}$.
- Margin $\frac{1}{\|w\|}$
- Constraints:
  - for $y_i = 1$ require $\mathbf{w}^\top \mathbf{x}_i + b \geq 1$
  - for $y_i = -1$ require $\mathbf{w}^\top \mathbf{x}_i + b \leq -1$

Hard margin SVM

- Binary classification problem with $y_i \in \{-1, 1\}$.
- Margin $\frac{1}{\|w\|}$
- Constraints:
  - for $y_i = 1$ require $\mathbf{w}^\top \mathbf{x}_i + b \geq 1$
  - for $y_i = -1$ require $\mathbf{w}^\top \mathbf{x}_i + b \leq -1$

This leads to

$$\boxed{\min \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad \forall i, \quad y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1}$$

Hard margin SVM

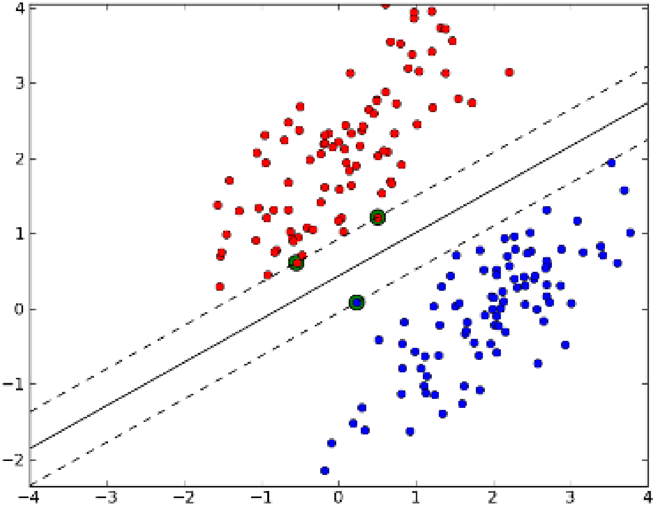
- Binary classification problem with $y_i \in \{-1, 1\}$.
- Margin $\frac{1}{\|w\|}$
- Constraints:
  - for $y_i = 1$ require $\mathbf{w}^\top \mathbf{x}_i + b \geq 1$
  - for $y_i = -1$ require $\mathbf{w}^\top \mathbf{x}_i + b \leq -1$

This leads to

$$\boxed{\min \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad \forall i, \quad y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1}$$

- quadratic program (not a so useful property nowadays)
- unfeasible if the data is not separable

Hard-margin SVM



Soft margin SVM

- Authorize some points to be on the wrong side of the margin
- Penalize by a cost proportional to the distance to the margin
- Introduce some slack variables ξ_i measuring the violation for each datapoint.

Soft margin SVM

- Authorize some points to be on the wrong side of the margin
- Penalize by a cost proportional to the distance to the margin
- Introduce some slack variables ξ_i measuring the violation for each datapoint.

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

$$\text{s.t.} \quad \forall i, \begin{cases} y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Lagrangian of the SVM

$$\mathcal{L}(w, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\nu})$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\big(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big) - \sum_{i=1}^{n}\nu_i\xi_i$$

Lagrangian of the SVM

$$\mathcal{L}(w, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\nu})$$

$$\begin{aligned} &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i(\mathbf{w}^\top\mathbf{x}_i + b)\right) - \sum_{i=1}^{n}\nu_i\xi_i \\ &= \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}^\top\left(\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i\right) + \sum_{i=1}^{n}\xi_i\left(C - \alpha_i - \nu_i\right) - \sum_{i=1}^{n}\alpha_i y_i b + \sum_{i=1}^{n}\alpha_i \end{aligned}$$

Lagrangian of the SVM

$$\mathcal{L}(w, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\nu})$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\big(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big) - \sum_{i=1}^{n}\nu_i\xi_i$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}^\top\Big(\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i\Big) + \sum_{i=1}^{n}\xi_i\,(C - \alpha_i - \nu_i) - \sum_{i=1}^{n}\alpha_i y_i b + \sum_{i=1}^{n}\alpha_i$$

Stationarity of the Lagrangian

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \quad \frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \nu_i \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n}\alpha_i y_i.$$

Lagrangian of the SVM

$\mathcal{L}(w, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\nu})$

$$= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\big(1 - \xi_i - y_i(\mathbf{w}^\top\mathbf{x}_i + b)\big) - \sum_{i=1}^{n}\nu_i\xi_i$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}^\top\Big(\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i\Big) + \sum_{i=1}^{n}\xi_i\,(C - \alpha_i - \nu_i) - \sum_{i=1}^{n}\alpha_i y_i b + \sum_{i=1}^{n}\alpha_i$$

Stationarity of the Lagrangian

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \quad \frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \nu_i \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n}\alpha_i y_i.$$

So that $\nabla\mathcal{L} = 0$ leads to

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \quad 0 \le \alpha_i \le C \quad \text{and} \quad \sum_{i=1}^{n}\alpha_i y_i = 0.$$

Dual of the SVM

$$\max_{\boldsymbol{\alpha}} \; -\frac{1}{2}\Big\|\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i\Big\|^2 + \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \qquad \forall i,\; 0 \le \alpha_i \le C.$$

$$\max_{\boldsymbol{\alpha}} \; -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{D_y K D_y}\boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{1}$$

$$\text{s.t.} \quad \boldsymbol{\alpha}^\top \mathbf{y} = 0, \qquad 0 \le \boldsymbol{\alpha} \le C.$$

with

- $\mathbf{y}^\top = (y_1, \ldots, y_n)$ the vector of labels
- $\mathbf{D_y} = \mathrm{Diag}(\mathbf{y})$ a diagonal matrix with the label
- $\mathbf{K}$ the Gram matrix with $\mathbf{K}_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$

Dual of the SVM

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\sum_{1\leq i,j\leq n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad \forall i,\ 0 \leq \alpha_i \leq C.$$

Dual of the SVM

$$\max_{\boldsymbol{\alpha}} \; -\frac{1}{2}\Big\|\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i\Big\|^2 + \sum_{i=1}^{n} \alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad \forall i,\; 0 \le \alpha_i \le C.$$

$$\max_{\boldsymbol{\alpha}} \; -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{D_y K D_y} \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{1}$$

$$\text{s.t.} \quad \boldsymbol{\alpha}^\top \mathbf{y} = 0, \qquad 0 \le \boldsymbol{\alpha} \le C.$$

with

- $\mathbf{y}^\top = (y_1, \ldots, y_n)$ the vector of labels
- $\mathbf{D_y} = \mathrm{Diag}(\mathbf{y})$ a diagonal matrix with the label
- $\mathbf{K}$ the Gram matrix with $\mathbf{K}_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$

KKT conditions for the SVM

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$

$$\alpha_i + \nu_i = C \quad \text{(LS)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$

$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$

$$\xi_i \geq 0 \quad \text{(PF)}$$

$$\alpha_i \geq 0 \quad \text{(DF)}$$

$$\nu_i \geq 0 \quad \text{(DF)}$$

$$\alpha_i \big(1 - \xi_i - y_i f(\mathbf{x}_i)\big) = 0 \quad \text{(CS)}$$

$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$

$$\alpha_i + \nu_i = C \quad \text{(LS)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$

$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$

$$\xi_i \geq 0 \quad \text{(PF)}$$

$$\alpha_i \geq 0 \quad \text{(DF)}$$

$$\nu_i \geq 0 \quad \text{(DF)}$$

$$\alpha_i \big(1 - \xi_i - y_i f(\mathbf{x}_i)\big) = 0 \quad \text{(CS)}$$

$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

Let

- $I = \{i \mid \xi_i > 0\}$

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$
$$\alpha_i + \nu_i = C \quad \text{(LS)}$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$
$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$
$$\xi_i \geq 0 \quad \text{(PF)}$$
$$\alpha_i \geq 0 \quad \text{(DF)}$$
$$\nu_i \geq 0 \quad \text{(DF)}$$
$$\alpha_i \big(1 - \xi_i - y_i f(\mathbf{x}_i)\big) = 0 \quad \text{(CS)}$$
$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

Let

- $I = \{i \mid \xi_i > 0\}$
- $M = \{i \mid y_i f(\mathbf{x}_i) = 1\}$

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$

$$\alpha_i + \nu_i = C \quad \text{(LS)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$

$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$

$$\xi_i \geq 0 \quad \text{(PF)}$$

$$\alpha_i \geq 0 \quad \text{(DF)}$$

$$\nu_i \geq 0 \quad \text{(DF)}$$

$$\alpha_i \big(1 - \xi_i - y_i f(\mathbf{x}_i)\big) = 0 \quad \text{(CS)}$$

$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

Let

- $I = \{i \mid \xi_i > 0\}$
- $M = \{i \mid y_i f(\mathbf{x}_i) = 1\}$
- $S = \{i \mid \alpha_i \neq 0\}$

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$

$$\alpha_i + \nu_i = C \quad \text{(LS)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$

$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$

$$\xi_i \geq 0 \quad \text{(PF)}$$

$$\alpha_i \geq 0 \quad \text{(DF)}$$

$$\nu_i \geq 0 \quad \text{(DF)}$$

$$\alpha_i \big(1 - \xi_i - y_i f(\mathbf{x}_i)\big) = 0 \quad \text{(CS)}$$

$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

Let

- $I = \{i \mid \xi_i > 0\}$
- $M = \{i \mid y_i f(\mathbf{x}_i) = 1\}$
- $S = \{i \mid \alpha_i \neq 0\}$
- $W = (I \cup M)^c$

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$

$$\alpha_i + \nu_i = C \quad \text{(LS)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$

$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$

$$\xi_i \geq 0 \quad \text{(PF)}$$

$$\alpha_i \geq 0 \quad \text{(DF)}$$

$$\nu_i \geq 0 \quad \text{(DF)}$$

$$\alpha_i \big(1 - \xi_i - y_i f(\mathbf{x}_i)\big) = 0 \quad \text{(CS)}$$

$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

Let

- $I = \{i \mid \xi_i > 0\}$
- $M = \{i \mid y_i f(\mathbf{x}_i) = 1\}$
- $S = \{i \mid \alpha_i \neq 0\}$
- $W = (I \cup M)^c$

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$

$$\alpha_i + \nu_i = C \quad \text{(LS)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$

$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$

$$\xi_i \geq 0 \quad \text{(PF)}$$

$$\alpha_i \geq 0 \quad \text{(DF)}$$

$$\nu_i \geq 0 \quad \text{(DF)}$$

$$\alpha_i \big(1 - \xi_i - y_i f(\mathbf{x}_i)\big) = 0 \quad \text{(CS)}$$

$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

Let

- $I = \{i \mid \xi_i > 0\}$
- $M = \{i \mid y_i f(\mathbf{x}_i) = 1\}$
- $S = \{i \mid \alpha_i \neq 0\}$
- $W = (I \cup M)^c$

$i \in I \Rightarrow \nu_i = 0 \Rightarrow \alpha_i = C \Rightarrow i \in S$

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$

$$\alpha_i + \nu_i = C \quad \text{(LS)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$

$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$

$$\xi_i \geq 0 \quad \text{(PF)}$$

$$\alpha_i \geq 0 \quad \text{(DF)}$$

$$\nu_i \geq 0 \quad \text{(DF)}$$

$$\alpha_i \big(1 - \xi_i - y_i f(\mathbf{x}_i)\big) = 0 \quad \text{(CS)}$$

$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

Let

- $I = \{i \mid \xi_i > 0\}$
- $M = \{i \mid y_i f(\mathbf{x}_i) = 1\}$
- $S = \{i \mid \alpha_i \neq 0\}$
- $W = (I \cup M)^c$

$i \in I \Rightarrow \nu_i = 0 \Rightarrow \alpha_i = C \Rightarrow i \in S$

$i \in W \Rightarrow \alpha_i = 0 \Leftrightarrow i \notin S$

We have $0 \leq \alpha_i \leq C$.

KKT conditions for the SVM

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}$$

$$\alpha_i + \nu_i = C \quad \text{(LS)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}$$

$$1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}$$

$$\xi_i \geq 0 \quad \text{(PF)}$$

$$\alpha_i \geq 0 \quad \text{(DF)}$$

$$\nu_i \geq 0 \quad \text{(DF)}$$

$$\alpha_i (1 - \xi_i - y_i f(\mathbf{x}_i)) = 0 \quad \text{(CS)}$$

$$\nu_i \xi_i = 0 \quad \text{(CS)}$$

with $f(\mathbf{x}_i) = \mathbf{w}^\top x_i + b$

Let

- $I = \{i \mid \xi_i > 0\}$
- $M = \{i \mid y_i f(\mathbf{x}_i) = 1\}$
- $S = \{i \mid \alpha_i \neq 0\}$
- $W = (I \cup M)^c$

$i \in I \Rightarrow \nu_i = 0 \Rightarrow \alpha_i = C \Rightarrow i \in S$

$i \in W \Rightarrow \alpha_i = 0 \Leftrightarrow i \notin S$

We have $0 \leq \alpha_i \leq C$.

The set S of support vectors is therefore composed of some points on the margin and all incorrectly placed points.

SVM summary so far

- Optimization problem formulated as a strongly convex QP

SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP

SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight α_i

SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight α_i
- The optimal solution is $\mathbf{w}^* = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i$, i.e. a weighted combination of the support vectors

SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight α_i
- The optimal solution is $\mathbf{w}^* = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points

$\rightarrow$ Leads to working set strategies.

$\rightarrow$ Computational gain

SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight α_i
- The optimal solution is $\mathbf{w}^* = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points

→ Leads to working set strategies.

→ Computational gain

Remarks:

1. the dual solution $\boldsymbol{\alpha}^*$ is not necessarily unique ⇒ there might be several possible sets of support vectors.

SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight α_i
- The optimal solution is $\mathbf{w}^* = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points

→ Leads to working set strategies.

→ Computational gain

Remarks:

1. the dual solution $\boldsymbol{\alpha}^*$ is not necessarily unique $\Rightarrow$ there might be several possible sets of support vectors.
2. How do we determine b?

SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight α_i
- The optimal solution is $\mathbf{w}^* = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points

$\rightarrow$ Leads to working set strategies.

$\rightarrow$ Computational gain

Remarks:

1. the dual solution $\boldsymbol{\alpha}^*$ is not necessarily unique $\Rightarrow$ there might be several possible sets of support vectors.
2. How do we determine b?

Representer property for the SVM

$$
\begin{aligned}
f^*(\mathbf{x}) &= \mathbf{w}^{*\top}\mathbf{x} + b \\
&= \sum_{i \in S} \alpha_i^* y_i \, \mathbf{x}_i^\top \mathbf{x} + b \\
&= \sum_{i \in S} \alpha_i^* y_i \, k(\mathbf{x}_i, \mathbf{x}) + b
\end{aligned}
$$

with $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$.

Representer property for the SVM

$$
\begin{aligned}
f^*(\mathbf{x}) &= {\mathbf{w}^*}^\top \mathbf{x} + b \\
&= \sum_{i \in S} \alpha_i^* y_i \, \mathbf{x}_i^\top \mathbf{x} + b \\
&= \sum_{i \in S} \alpha_i^* y_i \, k(\mathbf{x}_i, \mathbf{x}) + b
\end{aligned}
$$

with $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$.

- Eventually, this whole formulation depends only on the dot product between points

Representer property for the SVM

$$\begin{aligned} f^*(\mathbf{x}) &= {\mathbf{w}^*}^\top \mathbf{x} + b \\ &= \sum_{i \in S} \alpha_i^* y_i \, \mathbf{x}_i^\top \mathbf{x} + b \\ &= \sum_{i \in S} \alpha_i^* y_i \, k(\mathbf{x}_i, \mathbf{x}) + b \end{aligned}$$

with $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$.

- Eventually, this whole formulation depends only on the dot product between points

$\rightarrow$ Can we use another dot product than the one associated to the usual Euclidean distance in $\mathbb{R}^p$?

Hinge loss interpretation of the SVM

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

$$\text{s.t.} \quad \forall i, \begin{cases} y_i\left(\mathbf{w}^\top\mathbf{x}_i + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Hinge loss interpretation of the SVM

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

$$\text{s.t.} \quad \forall i, \begin{cases} \xi_i \geq 1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \\ \xi_i \geq 0 \end{cases}$$

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\max\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right)$$

Define the hinge loss $\ell(a, y) = (1 - ya)_+ \quad$ with $\quad (u)_+ = \max(u, 0)$.

Hinge loss interpretation of the SVM

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

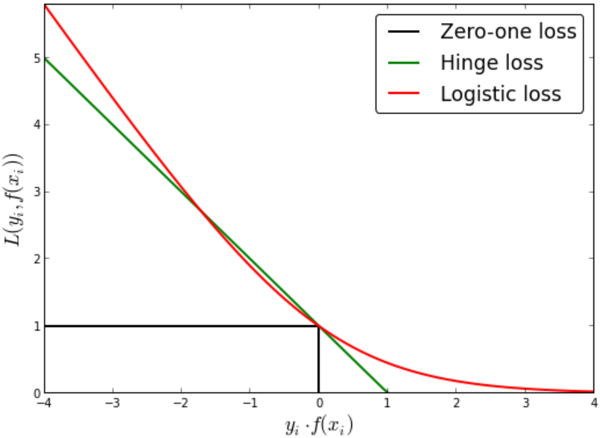
$$\text{s.t.} \quad \forall i, \begin{cases} y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\max\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right)$$

Define the hinge loss $\ell(a, y) = (1 - ya)_+ \quad \text{with} \quad (u)_+ = \max(u, 0)$.
Our problem is now of the form

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) + \frac{1}{2C}\|\mathbf{w}\|^2 \qquad \text{with} \quad f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b.$$

Hinge loss vs other losses



The hinge loss is the "least convex" loss which upper bounds the 0-1 loss and equals 0 for large scores.

SVM with the quadratic hinge loss

Quadratic hinge loss: $\ell(a, y) = (1 - ya)_+^2.$

SVM with the quadratic hinge loss

Quadratic hinge loss: $\ell(a, y) = (1 - ya)_+^2.$

Quadratic SVM

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right)^2$$

SVM with the quadratic hinge loss

Quadratic hinge loss: $\ell(a, y) = (1 - ya)_+^2.$

Quadratic SVM

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \max\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right)^2$$

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

$$\text{s.t.} \quad \forall i, \begin{cases} y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

SVM with the quadratic hinge loss

Quadratic hinge loss: $\ell(a, y) = (1 - ya)_+^2$.

Quadratic SVM

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max\left(1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right), 0\right)^2$$

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

$$\text{s.t.} \quad \forall i, \begin{cases} y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

→ Penalizes more strongly misclassified points

→ Less robust to outliers

→ Tends to be less sparse

→ Score in $[0, 1]$ for n large, interpretable as a probability.

Imbalanced classification

Imbalanced classification

Learn a binary classifier from (x_i, y_i) pairs with

$\mathcal{P} = \{i \mid y_i = 1\} \qquad \mathcal{N} = \{i \mid y_i = -1\}$,

$$n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \qquad \text{and with} \qquad n_+ \ll n_-.$$

Imbalanced classification

Learn a binary classifier from (x_i, y_i) pairs with

$\mathcal{P} = \{i \mid y_i = 1\} \qquad \mathcal{N} = \{i \mid y_i = -1\}$,

$$n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \qquad \text{and with} \qquad n_+ \ll n_-.$$

Problem: to minimize the number of mistakes the classifier learnt might classify all points as negatives.

Imbalanced classification

Learn a binary classifier from (x_i, y_i) pairs with

$$\mathcal{P} = \{i \mid y_i = 1\} \qquad \mathcal{N} = \{i \mid y_i = -1\},$$

$$n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \qquad \text{and with} \qquad n_+ \ll n_-.$$

Problem: to minimize the number of mistakes the classifier learnt might classify all points as negatives.

Some ways to address the issue

- Subsample the negatives, and learn an *ensemble* of classifiers.

Imbalanced classification

Learn a binary classifier from (x_i, y_i) pairs with

$$\mathcal{P} = \{i \mid y_i = 1\} \qquad \mathcal{N} = \{i \mid y_i = -1\},$$

$$n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \qquad \text{and with} \qquad n_+ \ll n_-.$$

Problem: to minimize the number of mistakes the classifier learnt might classify all points as negatives.

Some ways to address the issue

- Subsample the negatives, and learn an *ensemble* of classifiers.
- Introduce different costs for the positives and negatives

$$\min_{\mathbf{w}\in\mathbb{R}^p} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C_+\sum_{i\in\mathcal{P}}\xi_i + C_-\sum_{i\in\mathcal{N}}\xi_i$$

$$\text{s.t.} \quad \forall i, \quad y_i\left(\mathbf{w}^\top\mathbf{x}_i + b\right) \geq 1 - \xi_i$$

Imbalanced classification

Learn a binary classifier from (x_i, y_i) pairs with

$\mathcal{P} = \{i \mid y_i = 1\} \qquad \mathcal{N} = \{i \mid y_i = -1\}$,

$$n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \qquad \text{and with} \qquad n_+ \ll n_-.$$

Problem: to minimize the number of mistakes the classifier learnt might classify all points as negatives.

Some ways to address the issue

- Subsample the negatives, and learn an *ensemble* of classifiers.
- Introduce different costs for the positives and negatives

$$\min_{\mathbf{w} \in \mathbb{R}^p} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C_+ \sum_{i \in \mathcal{P}} \xi_i + C_- \sum_{i \in \mathcal{N}} \xi_i$$
$$\text{s.t.} \quad \forall i, \quad y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i$$

- Naive choice: $C_+ = C/n_+$ and $C_- = C/n_-$

  ⚠ Is suboptimal in theory and in practice !!

  $\rightarrow$ Better to search for the optimal hyperparameter pair (C_+, C_-).