SVM in Dual Form

Farzaneh Mirzazadeh
Borrowed most parts from Prof Dale Schuurmans notes and Prof David Helmbold slides

University of California, Santa Cruz

Winter' 17

Part 1: Background about constrained optimization

Background

Unconstrained optimization of a differentiable function

At the optimal point, the gradient must be zero

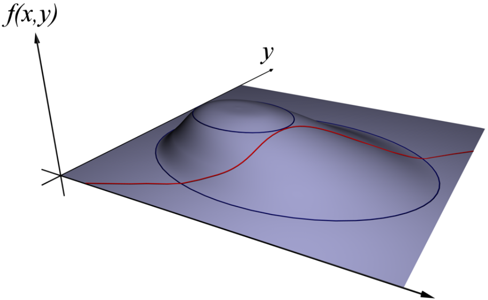
Background

Unconstrained optimization of a differentiable function

At the optimal point, the gradient must be zero

Optimization of a differentiable function with a single equality constraint

At the optimal point, the gradient of function must be parallel to the gradient of constraint.



Background: Constrained optimization

Note: Optional. Supposed to help in understanding the rest.

- When we have an unconstrained problem, we know how to find the local min: Just perform a local descent method.
- How about when we want to solve a constrained problem? How do we find the local optimum that satisfy the constraints?
- Example

$$\min_{x_1, x_2} \quad x_1 + x_2 \qquad \text{s.t.} \quad x_1^2 + x_2^2 = 2$$

Background: Constrained optimization

Note: Optional. Supposed to help in understanding the rest.

- When we have an unconstrained problem, we know how to find the local min: Just perform a local descent method.
- How about when we want to solve a constrained problem? How do we find the local optimum that satisfy the constraints?
- Example

$$\min_{x_1, x_2} \quad x_1 + x_2 \qquad \text{s.t.} \quad x_1^2 + x_2^2 = 2$$

- Find the points in which gradient of objective functions is parallel with the gradient of constraint set.
- Or equivalently (why?) form a function (Lagrangian)

$$\mathcal{L}(x_1, x_2, \lambda) = x_1 + x_2 + \lambda(x_1^2 + x_2^2 - 2)$$

$\nabla \mathcal{L} = \mathbf{0}$, solve for x_1, x_2, λ.

Lagrange Multipliers

Mathematical tool to transform a differentiable constrained optimization problems into optimization problems involving fewer constraints but more variables.

- How?
  1. Form a new function, Lagrangian, by additively combining the objective function and the constraint.
     Example from Last Slide:

     $$\mathcal{L}(x_1, x_2, \boldsymbol{\lambda}) = x_1 + x_2 - \boldsymbol{\lambda}(x_1^2 + x_2^2 - 2)$$

  2. Set the gradient of Lagrangian $\nabla\mathcal{L}$ to zero to find a stationary point of the $\mathcal{L}(x_1, x_2, \boldsymbol{\lambda})$.
  3. Solve for x_1, x_2, λ.
  4. Optimizers of the original problem form a stationary point (i.e.points with zero gradient) of the new problem.

Optimization with equality and inequality constraints

- The case where we have inequality constraints too.
- Form the generalized form of a Lagrangian as below.
- Satisfy Karush-KuhnTucker(KKT) conditions!

Primal

$\min_{\boldsymbol{w}} \ell(\boldsymbol{w}) \quad \text{s.t.} \quad A\boldsymbol{w} \geqslant b, C\boldsymbol{w} = \boldsymbol{d}$

Lagrangian with multiplier (λ) and multipliers (ν)

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \ell(\boldsymbol{w}) + \boldsymbol{\lambda}^\top(\boldsymbol{b} - A\boldsymbol{w}) + \boldsymbol{\nu}^\top(\boldsymbol{d} - C\boldsymbol{w})$$

KKT Conditions

1. Stationarity

$$\nabla \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}) == 0$$

2. Primal feasibility

$$A\boldsymbol{w} \geqslant b, C\boldsymbol{w} = \boldsymbol{d}$$

3. Dual feasibility

$$\boldsymbol{\lambda} \geqslant 0$$

4. Complementary slackness

$$\boldsymbol{\lambda}^\top (\boldsymbol{b} - A\boldsymbol{w}) = 0$$

   That is, for any single inequality constraint either the Lagrange multiplier is zero, or the inequality is tight.

For optimum of the constrained optimization problem, these conditions are satisfied.

Part 2: Deriving the dual for maximum margin classifier.

- Back to machine learning.
- Benefit of deriving dual of SVM
  - Gives another way of solving the optimization problem typically more efficient
  - Gives insight!
  - Again completely convex (Both primal and dual forms of max margin models are completely convex)

Recall: Maximum margin classification, the linearly separable case problem

Given a set of labeled examples, $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_t, y_t)$ where each $\boldsymbol{x}_i \in \mathbb{R}^n$ and each $y_i \in \{+1, -1\}$, find a weight vector $\boldsymbol{w}$ and intercept b such that $\mathrm{sign}(\boldsymbol{w} \bullet \boldsymbol{x}_i + b) = y_i$ for all i. (assume linearly separable)

Want to maximize the minimum *margin*, but

$$\max_{\boldsymbol{w}, b} \min_i y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i + b)$$

is not well defined (consider doubling $\boldsymbol{w}$ and b).

Recall: Maximum margin classification, the linearly separable case problem

Given a set of labeled examples, $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_t, y_t)$ where each $\boldsymbol{x}_i \in \mathbb{R}^n$ and each $y_i \in \{+1, -1\}$, find a weight vector $\boldsymbol{w}$ and intercept b such that $\mathrm{sign}(\boldsymbol{w} \bullet \boldsymbol{x}_i + b) = y_i$ for all i. (assume linearly separable)

Want to maximize the minimum *margin*, but

$$\max_{\boldsymbol{w},b} \min_i y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i + b)$$

is not well defined (consider doubling $\boldsymbol{w}$ and b).

functional margin $= y(\boldsymbol{w} \bullet \boldsymbol{x} + b)$ depends on scaling

geometric margin $=$ distance between point and hyperplane

$$= \frac{y(\boldsymbol{w} \bullet \boldsymbol{x} + b)}{\|\boldsymbol{w}\|_2}$$

Recall

Want to maximize geometric margin: $\min_i \frac{y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|_2}$

Equivalent to:

$$\min_{\boldsymbol{w},b} \|\boldsymbol{w}\|_2 \qquad \text{subject to} \qquad y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i + b) \geqslant 1 \text{ for all } i,$$

and to:

$$\min_{\boldsymbol{w},b} \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) \quad \text{subject to} \quad \Delta(\boldsymbol{y})\big(X\boldsymbol{w} + b\mathbf{1}\big) \geqslant \mathbf{1}$$

Primal Problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) \quad \text{subject to} \quad \Delta(\boldsymbol{y})\big(X\boldsymbol{w} + b\mathbf{1}\big) \geqslant \mathbf{1}$$

Primal Problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) \quad \text{subject to} \quad \Delta(\boldsymbol{y})\big(X\boldsymbol{w} + b\mathbf{1}\big) \geqslant \mathbf{1}$$

Lagrangian:

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) + \boldsymbol{\lambda}^\top\Big(\mathbf{1} - \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big)$$

Primal Problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) \quad \text{subject to} \quad \Delta(\boldsymbol{y})\big(X\boldsymbol{w} + b\mathbf{1}\big) \geqslant \mathbf{1}$$

Lagrangian:

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) + \boldsymbol{\lambda}^\top \Big(\mathbf{1} - \ \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big)$$

Dual problem:

$$\max_{\boldsymbol{\lambda}\succeq\mathbf{0}} \ \min_{\boldsymbol{w},b} \ \left[\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) + \boldsymbol{\lambda}^\top \Big(\mathbf{1} - \ \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big)\right]$$

Can solve this problem instead of the primal.

$$\max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \min_{\boldsymbol{w}, b} \quad \underbrace{\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) + \boldsymbol{\lambda}^\top \Big(\mathbf{1} - \quad \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big)}_{L(\boldsymbol{w}, b, \boldsymbol{\lambda})}$$

$$\max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \min_{\boldsymbol{w}, b} \quad \underbrace{\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) + \boldsymbol{\lambda}^\top \Big(\mathbf{1} - \quad \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big)}_{L(\boldsymbol{w}, b, \boldsymbol{\lambda})}$$

To solve inner min, differentiate $L(\boldsymbol{w}, b, \boldsymbol{\lambda})$ with respect to $\boldsymbol{w}$ and b, and set it to zero:

$$\max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \min_{\boldsymbol{w}, b} \quad \underbrace{\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) + \boldsymbol{\lambda}^\top \Big(\mathbf{1} - \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big)}_{L(\boldsymbol{w}, b, \boldsymbol{\lambda})}$$

To solve inner min, differentiate $L(\boldsymbol{w}, b, \boldsymbol{\lambda})$ with respect to $\boldsymbol{w}$ and b, and set it to zero:

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\lambda})}{\partial \boldsymbol{w}} = \boldsymbol{w} - X^\top \Delta(\boldsymbol{y})\boldsymbol{\lambda} = 0 \qquad \Rightarrow \; \boldsymbol{w} = X^\top \Delta(\boldsymbol{y})\boldsymbol{\lambda}$$

$$\max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \min_{\boldsymbol{w}, b} \underbrace{\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{w}) + \boldsymbol{\lambda}^\top \Big(\mathbf{1} - \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big)}_{L(\boldsymbol{w}, b, \boldsymbol{\lambda})}$$

To solve inner min, differentiate $L(\boldsymbol{w}, b, \boldsymbol{\lambda})$ with respect to $\boldsymbol{w}$ and b, and set it to zero:

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\lambda})}{\partial \boldsymbol{w}} = \boldsymbol{w} - X^\top \Delta(\boldsymbol{y})\boldsymbol{\lambda} = 0 \qquad \Rightarrow \quad \boldsymbol{w} = X^\top \Delta(\boldsymbol{y})\boldsymbol{\lambda}$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\lambda})}{\partial b} = -\boldsymbol{\lambda}^\top \boldsymbol{y} = 0 \qquad \Rightarrow \quad \boldsymbol{\lambda}^\top \boldsymbol{y} = 0$$

Interesting observations!

- $\boldsymbol{w} = X^\top \Delta(\boldsymbol{y})\boldsymbol{\lambda}$ means $\boldsymbol{w}$ is a weighted sum of examples (like what we had in representer theorem)

- $\boldsymbol{\lambda}^\top \boldsymbol{y} = 0$ means positive and negative examples have same total weight

- One of Karush-Kuhn-Tucker conditions, Complementary Slackness, implies that for each constraint term

$$\lambda_i \left(1 - y_i(\boldsymbol{w} \bullet \boldsymbol{x}_i + b)\right)$$

  if $\lambda_i \neq 0$ then the constraint is tight (i.e. $y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b) = 1$), so ...

- $\lambda_i > 0$ only when $\boldsymbol{x}_i$ is a support vector.
- $\boldsymbol{w}$ is a weighted sum of (signed) *support vectors*.

Get ready to plug into $L(\boldsymbol{w}, b, \boldsymbol{\lambda})$:

Term 1

$$\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} = \frac{1}{2}\underbrace{\left(X^\top\Delta(\boldsymbol{y})\boldsymbol{\lambda}\right)^\top}_{\boldsymbol{w}^\top}\underbrace{\left(X^\top\Delta(\boldsymbol{y})\boldsymbol{\lambda}\right)}_{\boldsymbol{w}} = \frac{1}{2}\boldsymbol{\lambda}^\top\Delta(\boldsymbol{y})XX^\top\Delta(\boldsymbol{y})\boldsymbol{\lambda}$$

Term 2

$$\begin{aligned}\boldsymbol{\lambda}^\top\Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1}) &= \boldsymbol{\lambda}^\top\Delta(\boldsymbol{y})X\underbrace{X^\top\Delta(\boldsymbol{y})\boldsymbol{\lambda}}_{\boldsymbol{w}} + \underbrace{b\boldsymbol{\lambda}^\top\boldsymbol{y}}_{0} \\ &= \boldsymbol{\lambda}^\top\Delta(\boldsymbol{y})XX^\top\Delta(\boldsymbol{y})\boldsymbol{\lambda}\end{aligned}$$

(Note $\Delta(\boldsymbol{y})\mathbf{1} = \boldsymbol{y}$)

The Dual

$$\max_{\lambda:\lambda \geqslant 0, \quad \lambda^\top y = 0} \quad \lambda^\top 1 - \frac{1}{2}\lambda^\top \Delta(y) X X^\top \Delta(y) \lambda$$

This is a quadratic programming problem - can be done numerically.

Recover primal solutions from λ^*

From λ^*, compute $\boldsymbol{w}^* = X^\top \Delta(\boldsymbol{y}) \lambda^*$, and
From complementary slackness for any $\lambda_i > 0$

Prediction

$$\hat{\boldsymbol{y}} = \mathrm{sign}(x_\circ^\top \boldsymbol{w}^* + b) = \mathrm{sign}(x_\circ^\top X^\top \Delta(y) \lambda + b)$$

Training and prediction kernelized!

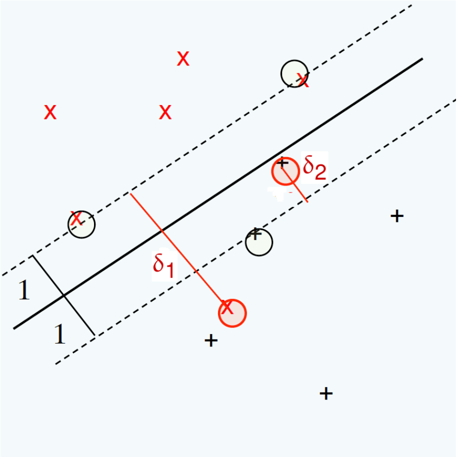
Sparseness

- Only for support vectors are λ_i non-zero – usually few support vectors.
- Removing labeled examples only changes hypothesis if a support vector removed.
- If ℓ out of t examples are support vectors, gives an expected error bound of ℓ/t.

Part 3: Deriving the dual for soft maximum margin classifier.

Recall: Soft margin Idea

- Data doesn't always have good margin
- Allow Margin errors (imperfect classification)
- Let $\delta_i \geqslant 0$ be error on $\boldsymbol{x}_i$
- *Hinge loss* is 0 when margin = 1, increases linearly as margin drops
- trade off accuracy and sum of "errors"



Deriving dual of soft maximum margin classification

Primal

$$\min_{\boldsymbol{w},b,\delta} \frac{\beta}{2}\|\boldsymbol{w}\|_2^2 + \mathbf{1}^\top\delta$$

s.t.

$$\delta \geqslant \mathbf{1} - \Delta(\boldsymbol{y})(X\boldsymbol{w} + \mathbf{1}b)$$

$$\delta \geqslant \mathbf{0}$$

Lagrangian with multiplier vectors λ, μ

$$L(\boldsymbol{w}, b, \delta, \lambda, \mu) = \frac{\beta}{2}\boldsymbol{w}^\top\boldsymbol{w} + \mathbf{1}^\top\delta + \lambda^\top\Big(\mathbf{1} - \delta - \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big) - \mu^\top\delta$$

Deriving Dual (Part I: Eliminating δ)

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, b, \delta, \lambda, \mu) =$$

$$\frac{\beta}{2}\boldsymbol{w}^\top\boldsymbol{w} \quad + \quad \mathbf{1}^\top\delta + \boldsymbol{\lambda}^\top\Big(\mathbf{1} - \delta - \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\Big) - \mu^\top\delta$$

From KKT Conditions: $\quad \boldsymbol{\lambda} \geqslant 0, \mu \geqslant 0$

Deriving Dual (Part I: Eliminating δ)

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, b, \delta, \lambda, \mu) =$$

$$\frac{\beta}{2}\boldsymbol{w}^\top\boldsymbol{w} \quad + \quad \mathbf{1}^\top\delta + \lambda^\top\left(\mathbf{1} - \delta - \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\right) - \mu^\top\delta$$

From KKT Conditions: $\boldsymbol{\lambda} \geqslant 0, \mu \geqslant 0$

- Compute the gradient of Lagrangian wrt variable that you want to eliminate and set it to zero.

$$\frac{dL}{d\delta} = \mathbf{1} - \boldsymbol{\lambda} - \mu = \mathbf{0} \implies \mathbf{1} = \boldsymbol{\lambda} + \mu$$

Deriving Dual (Part I: Eliminating δ)

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, b, \delta, \lambda, \mu) =$$

$$\frac{\beta}{2}\boldsymbol{w}^\top \boldsymbol{w} \quad + \quad \mathbf{1}^\top \delta + \lambda^\top \left(\mathbf{1} - \delta - \Delta(\boldsymbol{y})(X\boldsymbol{w} + b\mathbf{1})\right) - \mu^\top \delta$$

From KKT Conditions: $\boldsymbol{\lambda} \geqslant 0, \mu \geqslant 0$

- Compute the gradient of Lagrangian wrt variable that you want to eliminate and set it to zero.

$$\frac{dL}{d\delta} = \mathbf{1} - \boldsymbol{\lambda} - \boldsymbol{\mu} = \mathbf{0} \implies \mathbf{1} = \boldsymbol{\lambda} + \boldsymbol{\mu}$$

$$\implies \boldsymbol{\lambda} \leqslant \mathbf{1} \quad (\text{since} \quad \boldsymbol{\mu} \geqslant \mathbf{0}) \quad \text{Keep this in mind for later slides}$$

Deriving Dual (Part I: Eliminating δ)

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\mathbf{w}, b, \delta, \lambda, \mu) =$$

$$\frac{\beta}{2}\mathbf{w}^\top\mathbf{w} \quad + \quad \mathbf{1}^\top\boldsymbol{\delta} + \boldsymbol{\lambda}^\top\Big(\mathbf{1} - \boldsymbol{\delta} - \Delta(\mathbf{y})(X\mathbf{w} + b\mathbf{1})\Big) - \mu^\top\boldsymbol{\delta}$$

From KKT Conditions: $\boldsymbol{\lambda} \geqslant 0, \mu \geqslant 0$

- Compute the gradient of Lagrangian wrt variable that you want to eliminate and set it to zero.

$$\frac{dL}{d\delta} = \mathbf{1} - \boldsymbol{\lambda} - \mu = \mathbf{0} \implies \mathbf{1} = \boldsymbol{\lambda} + \mu$$

$\implies \boldsymbol{\lambda} \leqslant \mathbf{1}$ (since $\mu \geqslant \mathbf{0}$) Keep this in mind for later slides

- Plug-in and re-write the Lagrangian after eliminating the most recent variable.

$$\implies L(\mathbf{w}, b, \lambda) = \frac{\beta}{2}\mathbf{w}^\top\mathbf{w} + \boldsymbol{\lambda}^\top(1 - \Delta(y)(X\mathbf{w} + \mathbf{1}b)), \quad 0 \geqslant \boldsymbol{\lambda} \geqslant 1$$

Deriving Dual (Part II: eliminating b)

Again a similar procedure

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, b, \lambda) = \frac{\beta}{2}\boldsymbol{w}^\top \boldsymbol{w} + \boldsymbol{\lambda}^\top \big((1 - \Delta(y)(X\boldsymbol{w} + \mathbf{1}b)\big), \quad 0 \geqslant \boldsymbol{\lambda} \geqslant 1$$

Deriving Dual (Part II: eliminating b)

Again a similar procedure

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, b, \lambda) = \frac{\beta}{2}\boldsymbol{w}^\top \boldsymbol{w} + \boldsymbol{\lambda}^\top \big((1 - \Delta(y)(X\boldsymbol{w} + \mathbf{1}b)\big), \quad 0 \geqslant \boldsymbol{\lambda} \geqslant 1$$

- Compute the gradient of Lagrangian wrt variable that you want to eliminate and set it to zero.

$$\frac{dL}{db} = \boldsymbol{\lambda}^\top \boldsymbol{y} = \mathbf{0}$$

Deriving Dual (Part II: eliminating b)

Again a similar procedure

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, b, \lambda) = \frac{\beta}{2}\boldsymbol{w}^\top \boldsymbol{w} + \boldsymbol{\lambda}^\top \big((1 - \Delta(y)(X\boldsymbol{w} + \mathbf{1}b)\big), \quad 0 \geqslant \boldsymbol{\lambda} \geqslant 1$$

- Compute the gradient of Lagrangian wrt variable that you want to eliminate and set it to zero.

$$\frac{dL}{db} = \boldsymbol{\lambda}^\top \boldsymbol{y} = \mathbf{0}$$

- Plug-in and write the Lagrangian after eliminating the most recent variable.

$$L(\boldsymbol{w}, \boldsymbol{\lambda}) = \frac{\beta}{2}\boldsymbol{w}^\top \boldsymbol{w} + \boldsymbol{\lambda}^\top (1 - \Delta(y)(X\boldsymbol{w})), \quad 0 \geqslant \boldsymbol{\lambda} \geqslant 1$$

Deriving dual (Part III: eliminating $\boldsymbol{w}$)

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, \boldsymbol{\lambda}) = \frac{\beta}{2}\boldsymbol{w}^\top \boldsymbol{w} + \boldsymbol{\lambda}^\top \Big(1 - \Delta(y)(X\boldsymbol{w})\Big), \quad 0 \leqslant \boldsymbol{\lambda} \leqslant 1$$

Deriving dual (Part III: eliminating $\boldsymbol{w}$)

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, \boldsymbol{\lambda}) = \frac{\beta}{2}\boldsymbol{w}^\top \boldsymbol{w} + \boldsymbol{\lambda}^\top \Big(1 - \Delta(y)(X\boldsymbol{w})\Big), \quad 0 \leqslant \boldsymbol{\lambda} \leqslant 1$$

- Compute the gradient of Lagrangian wrt variable that you want to eliminate and set it to zero.

$$\frac{dL}{d\boldsymbol{w}} = \beta \boldsymbol{w} - X^\top \Delta(\boldsymbol{y})\boldsymbol{\lambda} = \mathbf{0} \quad \implies \boldsymbol{w} = \frac{1}{\beta} X^\top \Delta(\boldsymbol{y})\boldsymbol{\lambda}$$

- Plug-in and write the Lagrangian after eliminating most recent var.

$$\implies \max_{\boldsymbol{\lambda}: 0 \geqslant \boldsymbol{\lambda} \geqslant 1, \boldsymbol{\lambda}^\top y = 0} \boldsymbol{\lambda}^\top \mathbf{1} - \frac{1}{2\beta} \boldsymbol{\lambda}^\top \Delta(\boldsymbol{y}) X X^\top \Delta(\boldsymbol{y}) \boldsymbol{\lambda}$$

Deriving dual (Part III: eliminating $\boldsymbol{w}$)

- Write the Lagrangian with all dual and remaining primal variables.

$$L(\boldsymbol{w}, \lambda) = \frac{\beta}{2}\boldsymbol{w}^\top \boldsymbol{w} + \lambda^\top \Big(1 - \Delta(y)(X\boldsymbol{w})\Big), \quad 0 \leqslant \lambda \leqslant 1$$

- Compute the gradient of Lagrangian wrt variable that you want to eliminate and set it to zero.

$$\frac{dL}{d\boldsymbol{w}} = \beta \boldsymbol{w} - X^\top \Delta(\boldsymbol{y})\lambda = \boldsymbol{0} \quad \Longrightarrow \quad \boldsymbol{w} = \frac{1}{\beta} X^\top \Delta(\boldsymbol{y})\lambda$$

- Plug-in and write the Lagrangian after eliminating most recent var.

$$\Longrightarrow \max_{\lambda : 0 \geqslant \lambda \geqslant 1, \lambda^\top y = 0} \lambda^\top \boldsymbol{1} - \frac{1}{2\beta}\lambda^\top \Delta(\boldsymbol{y}) X X^\top \Delta(\boldsymbol{y})\lambda$$

Dual

$$\max_{\lambda :\, 0 \leqslant \lambda \leqslant 1, \lambda^\top y = 0} \lambda^\top \boldsymbol{1} - \frac{1}{2\beta}\lambda^\top \Delta(y) X X^\top \Delta(\boldsymbol{y})\lambda$$

Recovering primal solutions from dual solutions

Recover $\boldsymbol{w}^*$

$$\boldsymbol{w}^* = \frac{1}{\beta} X^\top \Delta(\boldsymbol{y}) \boldsymbol{\lambda}^*$$

Recover b^*

From complementary slackness

For any λ_i s.t. $0 < \lambda_i < 1$ we will have:

1. Because $\lambda_i < 1$, and we had from KKT conditions that $\lambda_i + \mu_i = 1$ $\implies \mu_i > 0 \implies \delta_i = 0$.
2. Because $0 < \lambda_i$, $1 - \delta_i - y_i(X_{i:}\boldsymbol{w}^* + b^*) = 0$

$$1 = y_i(X_{i:}\boldsymbol{w}^* + b^*), \quad y_i \in \{-1, +1\} \quad \implies \frac{1}{y_i} = (X_{i:}\boldsymbol{w}^* + b^*)$$

For any $y_i \in \{-1, +1\}$, $y_i^2 = 1$ so $\frac{1}{y_i} = y_i \implies y_i = X_{i:}\boldsymbol{w}^* + b^*$

Pick λ_i^*, $0 < \lambda_i^* < 1$ then set $b^* = y_i - X_{i:}\boldsymbol{w}^*$

Classification

Given $\boldsymbol{x}_\circ$,

$$\begin{aligned}\hat{y} &= \mathrm{sign}(\boldsymbol{x}_\circ^\top \boldsymbol{w}^* + b)\\ &= \mathrm{sign}(\boldsymbol{x}_\circ^\top X^\top \Delta(\boldsymbol{y}) + b)\end{aligned}$$

Still kernelized

- Can do all training and testing using λ, $XX^\top$, $\boldsymbol{x}^\top X^\top$.
- Don't need $\boldsymbol{w}$, nor explicit feature representation X.

- Replace $XX^\top$ with any training kernel K.
- Replace $x_\circ X^\top$ with any test kernel $Ktest$.

Note: For the soft margin case, we already knew kernelization was possible. Since the maximum soft margin problem is equivalent to L_2 regularized hinge loss. And by representer theorem, $L2$ regularized losses could be kernelized.

Where to read?

- Boyd and Vanderberghe (2004), Sec 5.2.3, 5.3.2
- Hastie et al., 2nd ed (2009), Sec. 12.1, 12.2, 12.3.1, 12.3.5
- Bishop (2006), Sec 7.1
- Cherkassky & Mullier (1998), Sec 9.1-9.3