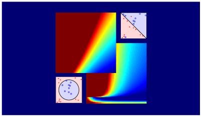
Machine Learning Techniques
(機器學習技巧)



Lecture 2: Dual Formulation of SVM

Hsuan-Tien Lin (林軒田)

htlin@csie.ntu.edu.tw

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)

Agenda

Lecture 2: Dual Formulation of SVM

- Motivation of Dual SVM
- Lagrange Dual SVM
- Solving Dual SVM
- Messages behind Dual SVM

Non-Linear Support Vector Machine Revisited

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{sub. to} \quad y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1,$$

$$\text{for } n = 1, 2, \ldots, N$$

Non-Linear Hard-Margin SVM

1. $\mathrm{A} = \begin{bmatrix} 0 & \mathbf{0}_{\tilde{d}}^T \\ \mathbf{0}_{\tilde{d}} & \mathrm{I}_{\tilde{d}} \end{bmatrix}$; $\mathbf{c} = \mathbf{0}_{\tilde{d}+1}$;
   $\mathbf{p}_n^T = y_n \begin{bmatrix} 1 & \mathbf{z}_n^T \end{bmatrix}$; $r_n = 1$
2. $\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \leftarrow \text{QP}(\mathrm{A}, \mathbf{c}, \mathrm{P}, \mathbf{r})$
3. return $b \in \mathbb{R}$ & $\mathbf{w} \in \mathbb{R}^{\tilde{d}}$ with
   $g_{\text{SVM}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{\Phi}(\mathbf{x}) + b)$

- demanded: **not many** (large-margin), but **sophisticated** boundary (feature transform)
- QP with $\tilde{d} + 1$ variables and N constraints
  —challenging if $\tilde{d}$ large, **or infinite?! :-)**

goal: SVM **without dependence on $\tilde{d}$**

Todo: SVM 'Without' $\tilde{d}$

Original SVM

(convex) QP of

- $\tilde{d} + 1$ variables
- N constraints

'Equivalent' SVM

(convex) QP of

- N variables
- $N + 1$ constraints

Warning: Heavy Math!!!!!!

- introduce some necessary math without rigor to help **understand SVM deeper**
- **'claim' some results** if details unnecessary
  —like how we 'claimed' Hoeffding

'Equivalent' SVM: based on some **dual problem** of Original SVM

Key Tool: Lagrange Multipliers

Regularization by Constrained-Minimizing E_{in}	$\Leftrightarrow$	Regularization by Minimizing E_{aug}
$\min_{\mathbf{w}} E_{in}(\mathbf{w})$ s.t. $\mathbf{w}^T\mathbf{w} \le C$		$\min_{\mathbf{w}} E_{aug}(\mathbf{w}) = E_{in}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$

- C equivalent to some $\lambda \ge 0$ by checking **optimality condition**

$$\nabla E_{in}(\mathbf{w}) + \frac{2\lambda}{N}\mathbf{w} = \mathbf{0}$$

- regularization: view λ as **given parameter instead of C**, and solve 'easily'
- dual SVM: view λ's as unknown given the constraints, and **solve them as variables instead**

how many λ's as variables?
N—one per constraint

Starting Point: Constrained to 'Unconstrained'

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

s.t. $\quad y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1,$

for $n = 1, 2, \ldots, N$

Lagrange Function

with Lagrange multipliers α_n,

$$\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) = \underbrace{\frac{1}{2}\mathbf{w}^T\mathbf{w}}_{\text{objective}} + \sum_{n=1}^{N} \alpha_n \underbrace{(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b))}_{\text{constraint}}$$

Claim

$$\text{SVM} \equiv \min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) = \min_{b,\mathbf{w}} \left(\infty \text{ if violate ; } \frac{1}{2}\mathbf{w}^T\mathbf{w} \text{ if feasible} \right)$$

- any 'violating' $(b, \mathbf{w})$: $\max_{\text{all } \alpha_n \geq 0} \square + \sum_n \alpha_n(\text{some positive}) \to \infty$
- any 'feasible' $(b, \mathbf{w})$: $\max_{\text{all } \alpha_n \geq 0} \square + \sum_n \alpha_n(\text{all non-positive}) = \square$

constraints now **hidden in** max

Fun Time

Lagrange Dual Problem

for any fixed $\boldsymbol{\alpha}'$ with all $\alpha'_n \geq 0$,

$$\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) \geq \min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}')$$

because max $\geq$ any

for best $\boldsymbol{\alpha}' \geq \mathbf{0}$ on RHS,

$$\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) \geq \underbrace{\max_{\text{all } \alpha_n' \geq 0} \min_{b,\mathbf{w}} \left(\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}') \right)}_{\text{Lagrange dual problem}}$$

because best is one of any

Lagrange dual problem:

'outer' maximization of $\boldsymbol{\alpha}$ on lower bound of original problem

Strong Duality of Quadratic Programming

$$\underbrace{\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right)}_{\text{equiv. to original (primal) SVM}} \geq \underbrace{\max_{\text{all } \alpha_n \geq 0} \left(\min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right)}_{\text{Lagrange dual}}$$

- '$\geq$': weak duality
- '$=$': **strong duality**, true for QP if
  - convex primal
  - feasible primal (true if $\mathbf{\Phi}$-separable)
  - linear constraints

—called constraint qualification

exists primal-dual optimal
solution $(b, \mathbf{w}, \boldsymbol{\alpha})$ for both sides

Solving Lagrange Dual: Simplifications (1/2)

$$\max_{\text{all } \alpha_n \ge 0} \left(\min_{b,\mathbf{w}} \underbrace{\frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n + b))}_{\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})} \right)$$

- inner problem 'unconstrained', at optimal: $\frac{\partial \mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}{\partial b} = 0 = -\sum_{n=1}^{N} \alpha_n y_n$
- no loss of optimality if solving with constraint $\sum_{n=1}^{N} \alpha_n y_n = 0$

but wait, b **can be removed**

$$\max_{\text{all } \alpha_n \ge 0, \sum y_n \alpha_n = 0} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n)) - \cancel{\sum_{n=1}^{N} \alpha_n y_n \cdot b} \right)$$

Solving Lagrange Dual: Simplifications (2/2)

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n)) \right)$$

- inner problem 'unconstrained', at optimal: $\frac{\partial \mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}{\partial w_i} = 0 = w_i - \sum_{n=1}^{N} \alpha_n y_n z_{n,i}$
- no loss of optimality if solving with constraint $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n$

but wait!

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n - \mathbf{w}^T\mathbf{w} \right)$$

$$\Longleftrightarrow \max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n$$

KKT Optimality Conditions

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\|\sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n\|^2 + \sum_{n=1}^{N} \alpha_n$$

if primal-dual optimal $(b, \mathbf{w}, \boldsymbol{\alpha})$,

- primal feasible: $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$
- dual-inner optimal: $\sum y_n \alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)) = 0$$

—called **Karush-Kuhn-Tucker (KKT) conditions**, necessary for optimality [& sufficient here]

will use **KKT** to 'solve' $(b, \mathbf{w})$ from optimal $\boldsymbol{\alpha}$

Fun Time

Dual Formulation of Support Vector Machine

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} \quad -\frac{1}{2} \left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n$$

standard hard-margin SVM dual

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} \alpha_n \\
\text{subject to} \quad & \sum_{n=1}^{N} y_n \alpha_n = 0; \\
& \alpha_n \geq 0, \text{ for } n = 1, 2, \ldots, N
\end{aligned}$$

(convex) QP of N variables & $N + 1$ constraints, as promised

how to solve? **yeah, we know QP! :-)**

Dual SVM with QP Solver

optimal $\boldsymbol{\alpha} = ?$

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} \alpha_n$$

subject to $\sum_{n=1}^{N} y_n \alpha_n = 0;$

$\alpha_n \geq 0,$

for $n = 1, 2, \ldots, N$

optimal $\boldsymbol{\alpha} \leftarrow \mathrm{QP}(\mathrm{A}, \mathbf{c}, \mathrm{P}, \mathbf{r})$

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{A} \boldsymbol{\alpha} + \mathbf{c}^T \boldsymbol{\alpha}$$

subject to $\mathbf{p}_m^T \boldsymbol{\alpha} \geq r_m,$

for $m = 1, 2, \ldots, M$

- $a_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m$
- $\mathbf{c} = -\mathbf{1}_N$
- $\mathbf{p}_{\geq} = \mathbf{y}$, $\mathbf{p}_{\leq} = -\mathbf{y}$;
  $\mathbf{p}_m^T = m$-th unit direction
- $r_{\geq} = 0$, $r_{\leq} = 0$; $r_m = 0$

note: many solvers treat **equality ($\mathbf{p}_{\geq}$, $\mathbf{p}_{\leq}$)** & **bound ($\mathbf{p}_m$)** constraints specially for **numerical stability**

Dual SVM with Special QP Solver

optimal $\alpha \leftarrow \text{QP}(\mathrm{Q}, \mathbf{c}, \mathrm{P}, \mathbf{r})$

$$\min_{\alpha} \quad \frac{1}{2}\alpha^T \mathrm{Q} \alpha + \mathbf{c}^T \alpha$$

subject to special equality and bound constraints

- $q_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m$, often non-zero
- if $N = 30,000$, dense Q (N by N symmetric) takes > 3G RAM
- need **special solver** for
  - not storing whole Q
  - utilizing special constraints properly

  to scale up to large N

usually better to use **special solver** in practice

Optimal $(b, \mathbf{w})$

KKT conditions

if primal-dual optimal $(b, \mathbf{w}, \boldsymbol{\alpha})$,

- primal feasible: $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$
- dual-inner optimal: $\sum y_n\alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)) = 0 \text{ (complementary slackness)}$$

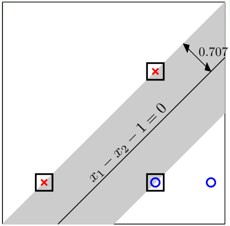
- optimal $\boldsymbol{\alpha} \Longrightarrow$ optimal $\mathbf{w}$? easy above!
- optimal $\boldsymbol{\alpha} \Longrightarrow$ optimal b? a range from primal feasible & equality from **comp. slackness** if one $\alpha_n > 0 \Rightarrow b = y_n - \mathbf{w}^T\mathbf{z}_n$

comp. slackness:
$\alpha_n > 0 \Rightarrow$ **on fat boundary (SV!)**

Fun Time

Support Vectors Revisited

- on boundary: 'locates' fattest hyperplane
  others: **not needed**
- examples with $\alpha_n > 0$: on boundary
- call $\alpha_n > 0$ examples $(\mathbf{z}_n, y_n)$
  support vectors ~~(candidates)~~
- SV (positive α_n)
  $\subseteq$ SV candidates (on boundary)



- only SV needed to compute $\mathbf{w}$: $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n = \sum_{\text{SV}} \alpha_n y_n \mathbf{z}_n$
- only SV needed to compute b: $b = y_n - \mathbf{w}^T \mathbf{z}_n$ with any SV $(\mathbf{z}_n, y_n)$

SVM: learn **fattest hyperplane**
by identifying **support vectors**
with **dual** optimal solution

Representation of Fattest Hyperplane

SVM

$$\mathbf{w}_{\text{SVM}} = \sum_{n=1}^{N} \alpha_n (y_n \mathbf{z}_n)$$

α_n from **dual solutions**

PLA

$$\mathbf{w}_{\text{PLA}} = \sum_{n=1}^{N} \beta_n (y_n \mathbf{z}_n)$$

β_n by **# mistake corrections**

$\mathbf{w}$ = linear combination of $y_n \mathbf{z}_n$

- also true for GD/SGD-based LogReg/LinReg when $\mathbf{w}_0 = \mathbf{0}$
- call $\mathbf{w}$ **'represented' by data**

SVM: represent $\mathbf{w}$ by SVs only

Summary: Two Forms of Hard-Margin SVM

Primal Hard-Margin SVM

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{sub. to} \quad y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1, \quad \text{for } n = 1, 2, \ldots, N$$

- $\tilde{d} + 1$ variables, N constraints —suitable when $\tilde{d} + 1$ small
- physical meaning: locate **specially-scaled** $(b, \mathbf{w})$

Dual Hard-Margin SVM

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{Q} \boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha}$$

$$\text{s.t.} \quad \mathbf{y}^T\boldsymbol{\alpha} = 0; \quad \alpha_n \geq 0 \text{ for } n = 1, \ldots, N$$

- N variables, $N + 1$ simple constraints —suitable when N small
- physical meaning: locate **SVs** $(\mathbf{z}_n, y_n)$ & their α_n

both eventually result in optimal $(b, \mathbf{w})$ for fattest hyperplane

$$g_{\text{SVM}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\boldsymbol{\Phi}(\mathbf{x}) + b)$$

Are We Done Yet?

goal: SVM **without dependence on $\tilde{d}$**

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{Q} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha}$$

$$\text{subject to} \quad \mathbf{y}^T \boldsymbol{\alpha} = 0;$$

$$\alpha_n \geq 0, \text{for } n = 1, 2, \ldots, N$$

- N variables, $N+1$ constraints: **no dependence on $\tilde{d}$?**
- $q_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m$: inner product in $\mathbb{R}^{\tilde{d}}$ —$O(\tilde{d})$ via naïve computation!

no dependence only if
avoiding naïve computation **(next lecture :-))**

Fun Time

Summary

Lecture 2: Dual Formulation of SVM

- Motivation of Dual SVM
  want to remove dependence on $\tilde{d}$
- Lagrange Dual SVM
  KKT conditions link primal/dual
- Solving Dual SVM
  another QP better solved with special solver
- Messages behind Dual SVM
  SVs represent fattest hyperplane