Lecture 2: Dual Support Vector Machine

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)
Roadmap

1. Embedding Numerous Features: Kernel Models
   - Lecture 1: Linear Support Vector Machine
     - Linear SVM: more robust and solvable with quadratic programming
   - Lecture 2: Dual Support Vector Machine
     - Motivation of Dual SVM
     - Lagrange Dual SVM
     - Solving Dual SVM
     - Messages behind Dual SVM

2. Combining Predictive Features: Aggregation Models
3. Distilling Implicit Features: Extraction Models
Dual Support Vector Machine

Motivation of Dual SVM

Non-Linear Support Vector Machine Revisited

\[ \min_{b,w} \frac{1}{2} w^T w \]

s. t. \[ y_n(w^T \phi(x_n) + b) \geq 1, \]

for \( n = 1, 2, \ldots, N \)

Non-Linear Hard-Margin SVM

1. \( Q = \begin{bmatrix} 0 & 0^T \tilde{d} \\ 0_{\tilde{d}} & I_{\tilde{d}} \end{bmatrix}; p = 0_{\tilde{d}+1} \)
   \( a_n^T = y_n \begin{bmatrix} 1 \\ z_n^T \end{bmatrix} \); \( c_n = 1 \)

2. \( \begin{bmatrix} b \\ w \end{bmatrix} \leftarrow \text{QP}(Q, p, A, c) \)

3. return \( b \in \mathbb{R} \) & \( w \in \mathbb{R}^{\tilde{d}} \) with
   \( g_{\text{SVM}}(x) = \text{sign}(w^T \phi(x) + b) \)

- 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 $E_{\text{in}}$

$$\min_w E_{\text{in}}(w) \text{ s.t. } w^T w \leq C$$

Regularization by Minimizing $E_{\text{aug}}$

$$\min_w E_{\text{aug}}(w) = E_{\text{in}}(w) + \frac{\lambda}{N} w^T w$$

- $C$ equivalent to some $\lambda \geq 0$ by checking optimality condition
  $$\nabla E_{\text{in}}(w) + \frac{2\lambda}{N} w = 0$$

- regularization: view $\lambda$ as given parameter instead of $C$, and solve ‘easily’

- dual SVM: view $\lambda$’s as unknown given the constraints, and solve them as variables instead

how many $\lambda$’s as variables?

$N$—one per constraint
Starting Point: Constrained to ‘Unconstrained’

\[ \begin{align*}
\min_{b, w} & \quad \frac{1}{2} w^T w \\
\text{s.t.} & \quad y_n(w^T z_n + b) \geq 1, \\
& \quad \text{for } n = 1, 2, \ldots, N
\end{align*} \]

Lagrange Function

with Lagrange multipliers \( \alpha_n \),

\[ L(b, w, \alpha) = \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n (1 - y_n(w^T z_n + b)) \]

Claim

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

- any ‘violating’ \((b, w)\): \( \max_{\alpha_n \geq 0} \left( \Box + \sum_n \alpha_n \text{(some positive)} \right) \rightarrow \infty \)
- any ‘feasible’ \((b, w)\): \( \max_{\alpha_n \geq 0} \left( \Box + \sum_n \alpha_n \text{(all non-positive)} \right) = \Box \)
Consider two transformed examples \((z_1, +1)\) and \((z_2, -1)\) with \(z_1 = z\) and \(z_2 = -z\). What is the Lagrange function \(L(b, w, \alpha)\) of hard-margin SVM?

1. \[ \frac{1}{2}w^T w + \alpha_1 (1 + w^T z + b) + \alpha_2 (1 + w^T z + b) \]
2. \[ \frac{1}{2}w^T w + \alpha_1 (1 - w^T z - b) + \alpha_2 (1 - w^T z + b) \]
3. \[ \frac{1}{2}w^T w + \alpha_1 (1 + w^T z + b) + \alpha_2 (1 + w^T z - b) \]
4. \[ \frac{1}{2}w^T w + \alpha_1 (1 - w^T z - b) + \alpha_2 (1 - w^T z - b) \]
Consider two transformed examples \((z_1, +1)\) and \((z_2, -1)\) with \(z_1 = z\) and \(z_2 = -z\). What is the Lagrange function \(L(b, w, \alpha)\) of hard-margin SVM?

1. \[
\frac{1}{2} w^T w + \alpha_1 (1 + w^T z + b) + \alpha_2 (1 + w^T z + b)
\]
2. \[
\frac{1}{2} w^T w + \alpha_1 (1 - w^T z - b) + \alpha_2 (1 - w^T z + b)
\]
3. \[
\frac{1}{2} w^T w + \alpha_1 (1 + w^T z + b) + \alpha_2 (1 + w^T z - b)
\]
4. \[
\frac{1}{2} w^T w + \alpha_1 (1 - w^T z - b) + \alpha_2 (1 - w^T z - b)
\]

Reference Answer: 2

By definition,

\[
L(b, w, \alpha) = \frac{1}{2} w^T w + \alpha_1 (1 - y_1 (w^T z_1 + b)) + \alpha_2 (1 - y_2 (w^T z_2 + b))
\]

with \((z_1, y_1) = (z, +1)\) and \((z_2, y_2) = (-z, -1)\).
for any fixed $\alpha'$ with all $\alpha'_n \geq 0$,

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

because $\max \geq \text{any}$

for best $\alpha' \geq 0$ on RHS,

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

Lagrange dual problem

because best is one of any

Lagrange dual problem:

‘outer’ maximization of $\alpha$ on lower bound of original problem
Strong Duality of Quadratic Programming

\[
\min_{b,w} \left( \max_{\alpha \geq 0} \mathcal{L}(b, w, \alpha) \right) \geq \max_{\alpha \geq 0} \left( \min_{b,w} \mathcal{L}(b, w, \alpha) \right)
\]

equiv. to original (primal) SVM

\[
\text{Lagrange dual}
\]

- ‘\(\geq\)’: weak duality
- ‘\(=\)’: strong duality, true for QP if
  - convex primal
  - feasible primal (true if \(\Phi\)-separable)
  - linear constraints

—called constraint qualification

exists primal-dual optimal solution \((b, w, \alpha)\) for both sides
Solving Lagrange Dual: Simplifications (1/2)

\[
\begin{aligned}
\max_{\alpha \geq 0} & \left( \min_{b,w} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n \left( 1 - y_n (w^T z_n + b) \right) \right) \\
\end{aligned}
\]

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

but wait, \( b \) can be removed

\[
\begin{aligned}
\max_{\alpha \geq 0, \sum y_n \alpha_n = 0} & \left( \min_{b,w} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n \left( 1 - y_n (w^T z_n) \right) - \sum_{n=1}^{N} \alpha_n y_n \cdot b \right) \\
\end{aligned}
\]
Solving Lagrange Dual: Simplifications (2/2)

\[
\max_{\alpha_n \geq 0, \sum y_n \alpha_n = 0} \left( \min_{b, w} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n (1 - y_n(w^T z_n)) \right)
\]

- **inner problem** ‘unconstrained’, at optimal:
  \[
  \frac{\partial \mathcal{L}(b, w, \alpha)}{\partial w_i} = 0 = w_i - \sum_{n=1}^{N} \alpha_n y_n z_n, i
  \]

- no loss of optimality if solving with constraint \( w = \sum_{n=1}^{N} \alpha_n y_n z_n \)

but wait!

\[
\max_{\alpha_n \geq 0, \sum y_n \alpha_n = 0, w=\sum \alpha_n y_n z_n} \left( \min_{b, w} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n - w^T w \right)
\]

\[\iff\]

\[
\max_{\alpha_n \geq 0, \sum y_n \alpha_n = 0, w=\sum \alpha_n y_n z_n} -\frac{1}{2} \left\| \sum_{n=1}^{N} \alpha_n y_n z_n \right\|^2 + \sum_{n=1}^{N} \alpha_n
\]
Dual Support Vector Machine

Lagrange Dual SVM

KKT Optimality Conditions

\[
\max_{\alpha_n \geq 0, \sum y_n \alpha_n = 0, w = \sum \alpha_n y_n z_n} \quad -\frac{1}{2} \left\| \sum_{n=1}^{N} \alpha_n y_n z_n \right\|^2 + \sum_{n=1}^{N} \alpha_n
\]

if primal-dual optimal \((b, w, \alpha)\),

- primal feasible: \(y_n(w^T z_n + b) \geq 1\)
- dual feasible: \(\alpha_n \geq 0\)
- dual-inner optimal: \(\sum y_n \alpha_n = 0; w = \sum \alpha_n y_n z_n\)
- primal-inner optimal (at optimal all ‘Lagrange terms’ disappear):
  \[
  \alpha_n (1 - y_n (w^T z_n + b)) = 0
  \]

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

will use KKT to ‘solve’ \((b, w)\) from optimal \(\alpha\)
For a single variable $w$, consider minimizing $\frac{1}{2}w^2$ subject to two linear constraints $w \geq 1$ and $w \leq 3$. We know that the Lagrange function $L(w, \alpha) = \frac{1}{2}w^2 + \alpha_1(1 - w) + \alpha_2(w - 3)$. Which of the following equations that contain $\alpha$ are among the KKT conditions of the optimization problem?

1. $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$
2. $w = \alpha_1 - \alpha_2$
3. $\alpha_1(1 - w) = 0$ and $\alpha_2(w - 3) = 0.$
4. all of the above
For a single variable $w$, consider minimizing $\frac{1}{2}w^2$ subject to two linear constraints $w \geq 1$ and $w \leq 3$. We know that the Lagrange function $\mathcal{L}(w, \alpha) = \frac{1}{2}w^2 + \alpha_1(1 - w) + \alpha_2(w - 3)$. Which of the following equations that contain $\alpha$ are among the KKT conditions of the optimization problem?

1. $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$
2. $w = \alpha_1 - \alpha_2$
3. $\alpha_1(1 - w) = 0$ and $\alpha_2(w - 3) = 0$.
4. all of the above

Reference Answer: 4

1. contains dual-feasible constraints;
2. contains dual-inner-optimal constraints;
3. contains primal-inner-optimal constraints.
**Dual Formulation of Support Vector Machine**

\[
\begin{align*}
\text{max} & \quad \sum_{n=1}^{N} \alpha_n y_n z_n - \frac{1}{2} \sum_{n=1}^{N} \alpha_n y_n z_n \\
\text{subject to} & \quad \sum_{n=1}^{N} y_n \alpha_n = 0; \\
& \quad \alpha_n \geq 0, \text{ for } n = 1, 2, \ldots, N
\end{align*}
\]

Standard hard-margin SVM dual

\[
\begin{align*}
\text{min} & \quad \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m z_n^T z_m - \sum_{n=1}^{N} \alpha_n \\
\text{subject to} & \quad \sum_{n=1}^{N} y_n \alpha_n = 0; \\
& \quad \alpha_n \geq 0, \text{ for } n = 1, 2, \ldots, N
\end{align*}
\]

(Convex) QP of \( N \) variables & \( N + 1 \) constraints, as promised

**how to solve? yeah, we know QP! :-)**
Dual SVM with QP Solver

optimal $\alpha = \text{?}$

\[
\begin{align*}
\min_{\alpha} & \quad \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m z_n^T z_m \\
- & \sum_{n=1}^{N} \alpha_n \\
\text{subject to} & \quad \sum_{n=1}^{N} y_n \alpha_n = 0; \\
& \quad \alpha_n \geq 0, \\
& \quad \text{for } n = 1, 2, \ldots, N
\end{align*}
\]

optimal $\alpha \leftarrow \text{QP}(Q, p, A, c)$

\[
\begin{align*}
\min_{\alpha} & \quad \frac{1}{2} \alpha^T Q \alpha + p^T \alpha \\
\text{subject to} & \quad a_i^T \alpha \geq c_i, \\
& \quad \text{for } i = 1, 2, \ldots \\
& \quad q_{n,m} = y_n y_m z_n^T z_m \\
& \quad p = -1_N \\
& \quad a_\geq = y, \quad a_\leq = -y; \\
& \quad a_n^T = n\text{-th unit direction} \\
& \quad c_\geq = 0, \quad c_\leq = 0; \quad c_n = 0
\end{align*}
\]

note: many solvers treat equality ($a_\geq, a_\leq$) & bound ($a_n$) constraints specially for numerical stability
Dual SVM with Special QP Solver

Optimal $\alpha \leftarrow \text{QP}(Q_D, p, A, c)$

$$\min_{\alpha} \frac{1}{2} \alpha^T Q_D \alpha + p^T \alpha$$

subject to special equality and bound constraints

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

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

If primal-dual optimal $(b, w, \alpha)$,

- **primal feasible**: $y_n(w^Tz_n + b) \geq 1$
- **dual feasible**: $\alpha_n \geq 0$
- **dual-inner optimal**: $\sum y_n \alpha_n = 0$; $w = \sum \alpha_n y_n z_n$
- **primal-inner optimal** (at optimal all ‘Lagrange terms’ disappear):

$$\alpha_n (1 - y_n(w^Tz_n + b)) = 0$$ (complementary slackness)

- optimal $\alpha \implies$ optimal $w$? easy above!
- optimal $\alpha \implies$ optimal $b$? a range from primal feasible & equality from comp. slackness if one $\alpha_n > 0 \implies b = y_n - w^Tz_n$

**comp. slackness:**

$$\alpha_n > 0 \implies \text{on fat boundary (SV!)}$$
Consider two transformed examples \((z_1, +1)\) and \((z_2, -1)\) with \(z_1 = z\) and \(z_2 = -z\). After solving the dual problem of hard-margin SVM, assume that the optimal \(\alpha_1\) and \(\alpha_2\) are both strictly positive. What is the optimal \(b\)?

1. \(-1\)
2. \(0\)
3. \(1\)
4. not certain with the descriptions above
Consider two transformed examples \((z_1, +1)\) and \((z_2, -1)\) with \(z_1 = z\) and \(z_2 = -z\). After solving the dual problem of hard-margin SVM, assume that the optimal \(\alpha_1\) and \(\alpha_2\) are both strictly positive. What is the optimal \(b\)?

\begin{itemize}
  \item[1] \(-1\)
  \item[2] \(0\)
  \item[3] \(1\)
  \item[4] not certain with the descriptions above
\end{itemize}

**Reference Answer:** 2

With the descriptions, at the optimal \((b, w)\),

\[ b = 1 - w^T z = -1 + w^T z \]

That is, \(w^T z = 1\) and \(b = 0\).
Support Vectors Revisited

- on boundary: ‘locates’ fattest hyperplane; others: not needed
- examples with $\alpha_n > 0$: on boundary
- call $\alpha_n > 0$ examples $(z_n, y_n)$ support vectors (candidates)
- SV (positive $\alpha_n$) $\subseteq$ SV candidates (on boundary)

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

SVM: learn fattest hyperplane by identifying support vectors with dual optimal solution
Dual Support Vector Machine

Messages behind Dual SVM

Representation of Fattest Hyperplane

SVM

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

\( \alpha_n \) from dual solution

PLA

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

\( \beta_n \) by # mistake corrections

\[ \mathbf{w} = \text{linear combination of } y_n \mathbf{z}_n \]

- also true for GD/SGD-based LogReg/LinReg when \( \mathbf{w}_0 = 0 \)
- call \( \mathbf{w} \) ‘represented’ by data

SVM: represent \( \mathbf{w} \) by SVs only
Summary: Two Forms of Hard-Margin SVM

**Primal Hard-Margin SVM**

\[
\begin{align*}
\min_{b,w} & \quad \frac{1}{2} w^T w \\
\text{sub. to} & \quad y_n(w^T z_n + b) \geq 1, \\
& \quad \text{for } n = 1, 2, \ldots, N
\end{align*}
\]

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

**Dual Hard-Margin SVM**

\[
\begin{align*}
\min_{\alpha} & \quad \frac{1}{2} \alpha^T Q_D \alpha - 1^T \alpha \\
\text{s.t.} & \quad y^T \alpha = 0; \\
& \quad \alpha_n \geq 0 \text{ for } n = 1, \ldots, N
\end{align*}
\]

- \(N\) variables, \(N + 1\) simple constraints
  —suitable when \(N\) small
- physical meaning: locate **SVs** \((z_n, y_n)\) & their \(\alpha_n\)

both eventually result in optimal \((b, w)\) for fattest hyperplane

\[g_{\text{SVM}}(x) = \text{sign}(w^T \Phi(x) + b)\]
Are We Done Yet?

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

$$\min_{\alpha} \frac{1}{2} \alpha^T Q_D \alpha - 1^T \alpha$$

subject to
$$y^T \alpha = 0;$$
$$\alpha_n \geq 0, \text{ for } n = 1, 2, \ldots, N$$

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

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

Consider applying dual hard-margin SVM on $N = 5566$ examples and getting 1126 SVs. Which of the following can be the number of examples that are on the fat boundary—that is, SV candidates?

1. 0
2. 1024
3. 1234
4. 9999

Reference Answer: 3

Because SVs are always on the fat boundary, $\# \text{SVs} \leq \# \text{SV candidates} \leq N$.
Consider applying dual hard-margin SVM on $N = 5566$ examples and getting 1126 SVs. Which of the following can be the number of examples that are on the fat boundary—that is, SV candidates?

1. 0
2. 1024
3. 1234
4. 9999

Reference Answer: 3

Because SVs are always on the fat boundary,

$$\# \text{ SVs} \leq \# \text{ SV candidates} \leq N.$$
Dual Support Vector Machine  

Summary

1 Embedding Numerous Features: Kernel Models

Lecture 2: Dual Support Vector Machine

- 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

- next: computing inner product in $\mathbb{R}^{\tilde{d}}$ efficiently

2 Combining Predictive Features: Aggregation Models

3 Distilling Implicit Features: Extraction Models