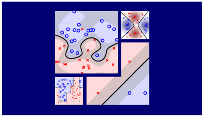
Machine Learning Techniques
(機器學習技法)



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

Non-Linear Support Vector Machine Revisited

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

$$\text{s. t.} \quad y_n(\mathbf{w}^T \underbrace{\mathbf{z}_n}_{\boldsymbol{\Phi}(\mathbf{x}_n)} + b) \geq 1,$$

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

Non-Linear Hard-Margin SVM

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

- 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}

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

$\Leftrightarrow$

Regularization by Minimizing E_{aug}

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

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

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

- 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'

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

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

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

Lagrange Function

with Lagrange multipliers $\cancel{\lambda_n}$ 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' $(b, \mathbf{w})$: $\max_{\text{all } \alpha_n \geq 0} \left(\square + \sum_n \alpha_n (\text{some positive}) \right) \rightarrow \infty$
- any 'feasible' $(b, \mathbf{w})$: $\max_{\text{all } \alpha_n \geq 0} \left(\square + \sum_n \alpha_n (\text{all non-positive}) \right) = \square$

constraints now **hidden in** max

Fun Time

Consider two transformed examples $(\mathbf{z}_1, +1)$ and $(\mathbf{z}_2, -1)$ with $\mathbf{z}_1 = \mathbf{z}$ and $\mathbf{z}_2 = -\mathbf{z}$. What is the Lagrange function $\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha})$ of hard-margin SVM?

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

Fun Time

Consider two transformed examples $(\mathbf{z}_1, +1)$ and $(\mathbf{z}_2, -1)$ with $\mathbf{z}_1 = \mathbf{z}$ and $\mathbf{z}_2 = -\mathbf{z}$. What is the Lagrange function $\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha})$ of hard-margin SVM?

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

Reference Answer: 2

By definition,

$$\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \alpha_1(1 - y_1(\mathbf{w}^T\mathbf{z}_1 + b)) + \alpha_2(1 - y_2(\mathbf{w}^T\mathbf{z}_2 + b))$$

with $(\mathbf{z}_1, y_1) = (\mathbf{z}, +1)$ and $(\mathbf{z}_2, y_2) = (-\mathbf{z}, -1)$.

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,

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

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

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

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

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 $\mathcal{L}(w, \boldsymbol{\alpha}) = \frac{1}{2}w^2 + \alpha_1(1 - w) + \alpha_2(w - 3)$. Which of the following equations that contain $\boldsymbol{\alpha}$ 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

Fun Time

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 $\mathcal{L}(w, \boldsymbol{\alpha}) = \frac{1}{2}w^2 + \alpha_1(1 - w) + \alpha_2(w - 3)$. Which of the following equations that contain $\boldsymbol{\alpha}$ 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

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

standard hard-margin SVM dual

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

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

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

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

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

Dual SVM with QP Solver

optimal $\boldsymbol{\alpha} = ?$

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

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

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

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

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

$$\text{subject to} \quad \mathbf{a}_i^T \boldsymbol{\alpha} \geq c_i, \quad \text{for } i = 1, 2, \ldots$$

- $q_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m$
- $\mathbf{p} = -\mathbf{1}_N$
- $\mathbf{a}_{\geq} = \mathbf{y}$, $\mathbf{a}_{\leq} = -\mathbf{y}$; $\mathbf{a}_n^T = n$-th unit direction
- $c_{\geq} = 0$, $c_{\leq} = 0$; $c_n = 0$

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

Dual SVM with Special QP Solver

optimal $\boldsymbol{\alpha} \leftarrow \text{QP}(\boxed{\mathrm{Q}_\mathrm{D}}, \mathbf{p}, \mathrm{A}, \mathbf{c})$

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \boxed{\mathrm{Q}_\mathrm{D}} \boldsymbol{\alpha} + \mathbf{p}^T \boldsymbol{\alpha}$$

subject to special equality and bound constraints

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

Consider two transformed examples $(\mathbf{z}_1, +1)$ and $(\mathbf{z}_2, -1)$ with $\mathbf{z}_1 = \mathbf{z}$ and $\mathbf{z}_2 = -\mathbf{z}$. After solving the dual problem of hard-margin SVM, assume that the optimal α_1 and α_2 are both strictly positive. What is the optimal b?

1. -1
2. 0
3. 1
4. not certain with the descriptions above

Fun Time

Consider two transformed examples $(\mathbf{z}_1, +1)$ and $(\mathbf{z}_2, -1)$ with $\mathbf{z}_1 = \mathbf{z}$ and $\mathbf{z}_2 = -\mathbf{z}$. After solving the dual problem of hard-margin SVM, assume that the optimal α_1 and α_2 are both strictly positive. What is the optimal b?

1. -1
2. 0
3. 1
4. not certain with the descriptions above

Reference Answer: 2

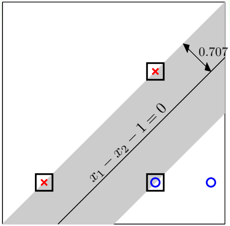
With the descriptions, at the optimal $(b, \mathbf{w})$,

$$b = +1 - \mathbf{w}^T\mathbf{z} = -1 + \mathbf{w}^T\mathbf{z}$$

That is, $\mathbf{w}^T\mathbf{z} = 1$ 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 $(\mathbf{z}_n, y_n)$ **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 solution**

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

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

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

Dual Hard-Margin SVM

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{Q_D} \boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha}$$
$$\text{s.t.} \quad \mathbf{y}^T\boldsymbol{\alpha} = 0;$$
$$\alpha_n \geq 0 \text{ for } n = 1, \ldots, N$$

- 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 $\tilde{d}$**

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{Q_D} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{y}^T \boldsymbol{\alpha} = 0; \\ & \alpha_n \geq 0, \text{for } n = 1, 2, \ldots, N \end{aligned}$$

- 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

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

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,

$$\# \text{ SVs} \leq \# \text{ SV candidates} \leq N.$$

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