DSA5105 LEC3 SVM

Summary

only works for Regression (not classification)

→ reformulate Ridge Reg as Kernel Ridge Reg.

↓

without specifizing feature map

SVM (here we use interchangably)

① Linear f : $f(x) = w^T x + b$ (actually is Affine function)

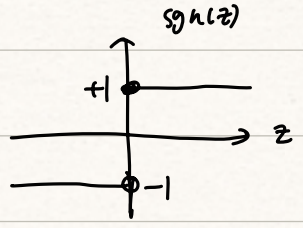
→ hyperplane ⇒ $f(x) = 0$

$\mathcal{H}_{svm} := \{ f : f(x) = \text{sign}(w^T x + b) \}$

② Binary Classification ← SVM

→ $\mathcal{D} = \{ (x_i, y_i) : x_i \in \mathbb{R}^d, \ y_i \in \{-1, +1\} \}_{i=1}^N$

→ linear decision function : $f(x) = \text{sgn}(w^T x + b)$



⇑

Linearly Separable Assumption :

$\exists f^*$ (linear decision f) s.t $f^*(x_i) = y_i$

$\Leftrightarrow \exists (\hat{w}, \hat{b})$ s.t $y_i(\hat{w}^T x_i + b_i) \geq 0$ for $i = 1, 2, \ldots, N$

③ There are many choices of 'Separable Hyperplane' (Non-unique)

↑

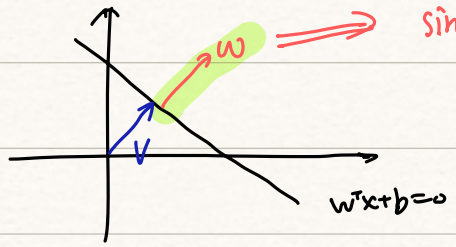
{ robust generalization (IDEA)

Maximize Margin

↳ to find the optimal hyperplane

Hyperplane

[E.g.]



$\Rightarrow$ since for $\forall x_1, x_2 \in H$, then $w^T(x_1 - x_2) = 0$

$\Rightarrow w \perp H$

How to calculate the distance from H to origin?

① $v \parallel w$

② $v \in H$

$\Rightarrow w^T v + b = 0$ & $v = \alpha w$

$\Rightarrow \alpha = -\frac{b}{\|w\|^2}$

$\Rightarrow$ distance $= \|v\| = \left\| -\frac{b}{\|w\|^2} w \right\| = \frac{|b|}{\|w\|}$

Proof

① $x - x_0 \parallel w$

② $x \in H \Rightarrow w^T x + b = 0$

③ distance $= \left| \left\langle x - x_0, \frac{w}{\|w\|} \right\rangle \right| = \frac{|w^T x_0 + b|}{\|w\|}$

generalization: distance $(x_0, H) = \frac{|w^T x_0 + b|}{\|w\|}$

---

④ Formulation

①
$$\max_{w,b} \ \gamma^g(w,b) \quad \text{s.t} \ (w,b) \text{ can separate correctly}$$

← the ideally formulation

②
$$\Leftrightarrow \max_{w,b} \min_{i=1,2\ldots N} \gamma_i^g(w,b) \longrightarrow \gamma_i^g = \frac{|w^T x_i + b|}{\|w\|}$$

$$\text{s.t} \quad y_i(w^T x_i + b) > 0 \quad \text{for all } i = 1,2,\ldots N$$

③
$$\Leftrightarrow \max_{w,b} \frac{1}{\|w\|} \min_{i=1,2\ldots N} |w^T x_i + b|$$

$$\text{s.t} \quad y_i(w^T x_i + b) > 0 \quad \forall i$$

equivalent

④
$$\Leftrightarrow \max_{w,b,\lambda} \frac{1}{\|w\|} \cdot \lambda$$

$$\text{s.t} \quad y_i(w^T x_i + b) \geq \lambda \quad \forall i$$

(The trick step)

⑤

⑤

$$\max_{w,b} \frac{1}{\|w\|} \quad \text{s.t} \quad y_i(w^T x_i + b) \geq 1 \quad \forall i$$

equivalent in the sense that it will share the same objective value but may have different optimizer

⑥

$$\min_{w,b} \frac{1}{2} w^T w \quad \text{s.t} \quad y_i(w^T x_i + b) \geq 1 \quad \forall i$$

Primal Form of SVM

2 Ques
- equivalence
- convexity

Dual Form (Lagrangian Multiplier)

↓ optimization problem

$$\min_{z} F(z) \quad \text{s.t} \quad z \in P \rightsquigarrow \begin{cases} g_i(x) = 0 \\ h_j(x) \leq 0 \end{cases}$$

Lagrangian Multiplier Framework (KKT necessary condition)

Lagrangian Function → $L(x; \lambda, \mu) = f(x) + \sum \lambda_i g_i(x) + \sum \mu_j h_j(x)$

KKT Condition [Necessary]

① Stationarity $\nabla_x L(\hat{x}; \hat{\lambda}, \hat{\mu}) = 0$

there exists $\hat{\lambda}, \hat{\mu}$, s.t $\hat{x} \in Z$ → Lagrangian Multipliers

② Dual feasibility $\hat{\mu} \geq 0$

③ Prime feasibility $g_i(\hat{x}) = 0$

$h_j(\hat{x}) \leq 0$

④ Complementary Slackness $\hat{\mu}_j h_j(\hat{x}) = 0$

↑ should under some constraint qualification condition

Application of KKT

i) distance between x_0 and H

$$\text{dist}(x_0, H) = \begin{cases} \min\limits_{x} \frac{1}{2}\|x - x_0\|_2^2 \\ \text{s.t} \quad w^T x + b = 0 \end{cases} \rightsquigarrow \text{convex problem}$$

$$L(x; \mu) = \frac{1}{2}\|x - x_0\|_2^2 + \mu(w^T x + b)$$

$$\nabla_x L(\hat{x}; \hat{\mu}) = 0 \Rightarrow (\hat{x} - x_0) + \hat{\mu} w = 0$$

$$\Rightarrow \boxed{w^T \hat{x}} - w^T x_0 + \hat{\mu} w^T w = 0$$

$\boxed{w^T \hat{x}} \downarrow -b$ → primary feasibility

$$\Rightarrow \hat{\mu} = \frac{b + w^T x_0}{\|w\|_2^2}$$

$$\Rightarrow \boxed{\text{dist}} = |\hat{x} - x_0| = \|\hat{\mu} w\|_2 = \frac{|w^T x_0 + b|}{\|w\|_2}$$

ii) Dual Form of SVM

Recall: Primal form

$$\begin{cases} \min\limits_{w} \frac{1}{2}\|w\|^2 \\ \text{s.t} \quad y_i(w^T x_i + b) \geq 1 \quad \text{for } \forall i \end{cases}$$

$$L(w, b; \mu) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \mu_i [1 - y_i(w^T x_i + b)]$$

Stationarity

① $\nabla_w L(\hat{w}, \hat{b}; \hat{\mu}) = 0 \Rightarrow \hat{w} = \sum_{i=1}^{N} \hat{\mu}_i y_i x_i$

$\nabla_b L(\hat{w}, \hat{b}, \hat{\mu}) = 0 \Rightarrow \sum \hat{\mu}_i y_i = 0$

② Dual Feasibility $\Rightarrow \hat{\mu}_i \geq 0 \quad \forall i = 1, \ldots, N$

③ Primal Feasibility $\Rightarrow y_i(\hat{w}^T x_i + \hat{b}) \geq 1 \quad \forall i = 1, \ldots, N$

④ Complementary Slackness $\Rightarrow \mu_i [1 - y_i(w^T x_i + b)] = 0 \quad \forall i = 1, 2, \ldots N$

Strong Duality guarantee that, these 2 are equivalent

$$\max_{\mu} \min_{w} L(w;\mu) \quad \text{s.t} \quad \mu \geq 0$$

→ for the convexity of $L(w;\mu)$

$$\min_w L(w;\mu) = L\left(\sum \mu_i y_i x_i ; \mu\right)$$

$$= -\frac{1}{2} \sum \mu_i \mu_j y_i y_j x_i^T x_j + \sum \mu_i$$

Dual Form

$$\longleftrightarrow \quad \max_{\mu} -\frac{1}{2} \sum \mu_i \mu_j y_i y_j x_i^T x_j + \sum \mu_i \quad \text{s.t} \quad \mu \geq 0$$

Suppose that we have $\hat{\mu} \geq 0$, then the decision function $\hat{f}$ becomes:

$$\hat{w} = \sum_{i=1}^{N} \hat{\mu}_i y_i x_i$$

$$\hat{f}(x) = \text{sign}\left(\hat{w}^T x + \hat{b}\right)$$

$$= \text{sign}\left(\sum_{i=1}^{N} \hat{\mu}_i y_i \langle x_i, x \rangle + \hat{b}\right)$$

$$= \text{sign}\left(\sum_{\alpha \in SV} \hat{\mu}_\alpha y_\alpha \langle x_\alpha, x \rangle + \hat{b}\right)$$

from CS condition, we have, some of $\hat{\mu}_i = 0$

$\alpha \in SV \Leftrightarrow \mu_\alpha > 0$ Support Vector

Sparse Solution

such data points tend to be smaller

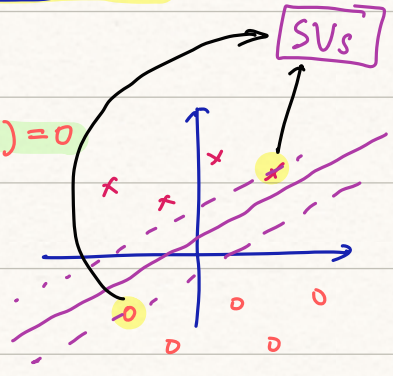
the summation is very sparse

advantage

Recall: CS condition → $\hat{\mu}_i (1 - y_i(\hat{w}^T x_i + b)) = 0$

for these vectors, they achieve:

$$y_\alpha (\hat{w}^T x_\alpha + \hat{b}) = 1$$



the distance to decision hyperplane $= \frac{1}{\|w\|}$

Kernel SVM

Recall:

$$\hat{f}(x) = \text{sgn}(\hat{w}^T x + \hat{b})$$

$$= \text{sgn}\left(\sum_i \hat{\mu}_i y_i \langle x_i, x \rangle + \hat{b}\right)$$

$\Downarrow$ only depends on $x_i^T x$ (or $\langle x_i, x \rangle$)

$\Downarrow$

IDEA: extend to Kernel trick!

(like Ridge Reg → Kernel Ridge Reg)

Punchline

deal with non-linearly separable dataset → Kernel SVM

$x \xrightarrow{\phi(\cdot)} \phi(x)$ → linearly separable in feature space

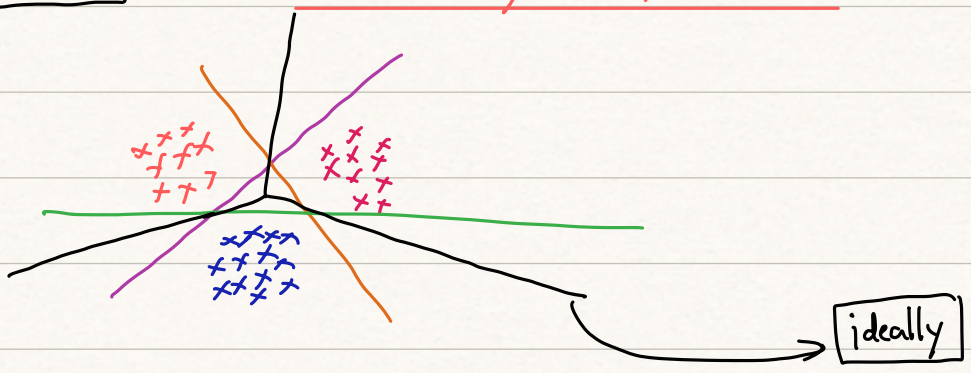
x: non-linearly separable

$$\hat{f}(x) = \text{sgn}\left(\sum_i \hat{\lambda}_i y_i k(x_i, x) + \hat{b} \right) \qquad k(x_i, x) = \langle \phi(x_i), \phi(x) \rangle$$

---

generalize to multi-class classification

1 vs all → several binary-classification tests



→ ideally

Q: How to aggregate these 3 sub-models

Recap of Lagrangian Duality

① KKT condition

given optimization problem:

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t} & g_i(x) \le 0 \quad i = 1, 2, \dots, I \\ & h_j(x) = 0 \quad j = 1, 2, \dots, J \end{cases}$$

(Necessary Part)

$\hat{x} \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x)$ s.t. $x \in C$

under some constraint qualification $\Longrightarrow$

① $\hat{x} \in C$

② $\exists \lambda_i \geq 0,\ \mu_i \in \mathbb{R}$ s.t.

$\nabla f(\hat{x}) + \sum_I \lambda_i \nabla g_i(x) + \sum_J \mu_j \nabla h_j(x) = 0$

$\lambda_i g_i(\hat{x}) = 0 \quad \forall i \in I$

Define $L(x;\lambda,\mu) = f(x) + \sum \lambda_i g_i(x) + \sum \mu_j h_j(x)$

$\Longleftrightarrow$

① $g_i(\hat{x}) \leq 0,\ h_j(\hat{x}) = 0$

② $\hat{\lambda}_i \geq 0,\ \hat{\mu}_i \in \mathbb{R}$

③ $\nabla_x L(\hat{x};\hat{\lambda},\hat{\mu}) = 0$

④ $\hat{\lambda}_i g_i(\hat{x}) = 0$

(Sufficient Part)

ⓐ $(\hat{x};\hat{\lambda},\hat{\mu})$ satisfy ① ② ③ ④

ⓑ Convex constraints for $\begin{cases} f(x) \\ g_i(x) \\ h_j(x) \end{cases}$ $\Longrightarrow$ $\hat{x} \in \underset{x \in \mathbb{R}}{\operatorname{argmin}} f(x)$ s.t. $x \in C$

Note: this part is not related to duality

KEY POINT

focus on the connection Between

- primal minimizer $\hat{x}$
- dual maximizer $\hat{\lambda}, \hat{\mu}$

② Duality Introduction

given $\begin{cases} \min\limits_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & g_i(x) \leq 0 \quad i = 1,2,\dots,I \\ & h_j(x) = 0 \quad j = 1,2,\dots,J \end{cases}$

Define $L(x;\lambda,\mu) = f(x) + \sum_{i \in I} \lambda_i g_i(x) + \sum \mu_j h_j(x)$

a)

then optimization problem $\Leftrightarrow$ $\min\limits_{x} \max\limits_{\substack{\lambda \geq 0 \\ \mu \in \mathbb{R}^J}} L(x;\lambda,\mu)$

b) Weak Duality ($\hat{x}$ → optimizer for constrained prob)

general result ← $\min_x L(x;\lambda,\mu) \le f(x)$ for { feasible x, $\lambda \ge 0$, μ }

$\Rightarrow \max_{\lambda \ge 0, \mu} \min_x L(x;\lambda,\mu) \le f(\hat{x}) := \min_{x \in C} f(x)$

dual ≤ PRIMAL

weaker slater: ① g_i affine, ② Feasible Region $\neq \phi$

slater: $\exists\, x \in \text{relint } \mathcal{D}$, s.t. $g_i(x) < 0$ & $h_j(x) = 0$

C) Strong Duality

① Slater's / Weak Slater's condition + { f convex, g_i convex, h_j affine } $\Rightarrow$ strong duality

② given strong duality holds, then

$(\hat{x};\hat{\lambda},\hat{\mu})$ is the optimal solution for $(P;D)$

$\Rightarrow$ {
① $g_i(\hat{x}) \le 0$ & $h_j(\hat{x}) = 0$ for $\forall i, j$
② $\hat{\lambda} \ge 0$
③ $\nabla_x L(\hat{x};\hat{\lambda},\hat{\mu}) = 0$
④ $\hat{\lambda}_i g_i(\hat{x}) = 0$ $\forall i = 1,2,\dots,I$
}

③ given strong duality holds, if (P) is a convex program { f convex, g_i convex, h_j affine }

(if slater's condition + convex program holds)

then $(\hat{x};\hat{\lambda},\hat{\mu})$ is the optimal solution for $(P;D)$

$\Leftrightarrow$ {
① $g_i(\hat{x}) \le 0$ & $h_j(\hat{x}) = 0$ for $\forall i, j$
② $\hat{\lambda} \ge 0$
③ $\nabla_x L(\hat{x};\hat{\lambda},\hat{\mu}) = 0$
④ $\hat{\lambda}_i g_i(\hat{x}) = 0$ $\forall i = 1,2,\dots,I$
}

④ **Saddle Point Defn**

$(\hat{x};\hat{\lambda},\hat{\mu})$ is saddle point of $L(x;\lambda,\mu)$

$$\Leftrightarrow \quad L(\hat{x};\hat{\lambda},\hat{\mu}) = \sup_{\substack{\lambda\geq 0\\ \mu}} \inf_{x} L(x;\lambda,\mu) = \inf_{x} \sup_{\substack{\lambda\geq 0\\ \mu}} L(x;\lambda,\mu)$$

Then, $(\hat{x};\hat{\lambda},\hat{\mu})$ is saddle point for $L(x;\lambda,\mu)$

$\Leftrightarrow$ strong duality holds & $(\hat{x};\hat{\lambda},\hat{\mu})$ is optimizer to $(P;D)$

SVM formulation: (Soft-Margin)

primal:
$$\begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2} w^T w + C \sum\limits_{i=1}^{N} \xi_i \\ \text{s.t} & y_i(w^T x_i + b) \geq 1 - \xi_i \qquad i = 1,2,\dots,N \\ & \xi_i \geq 0 \qquad\qquad\qquad i = 1,2,\dots,N \end{cases}$$

dual:
$$L(w,b,\xi;\alpha,\beta) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i (1 - \xi_i - y_i(w^T x_i + b)) - \sum_{i=1}^{N} \beta_i \xi_i$$

From KKT Condition, if $(\hat{w},\hat{b},\hat{\xi};\hat{\alpha},\hat{\beta})$ are optimizer to $(P;D)$

then it holds:

① $y_i(\hat{w}^T x_i + \hat{b}) \geq 1 - \hat{\xi}_i$

$\hat{\xi}_i \geq 0$

② $\hat{\alpha}_i \geq 0$, $\hat{\beta}_i \geq 0$

③ $\nabla_w L(\hat{w},\hat{b},\hat{\xi};\hat{\alpha},\hat{\beta}) = \hat{w} - \sum\limits_{i=1}^{N} \hat{\alpha}_i y_i x_i = 0$

$\nabla_b L(\hat{w},\hat{b},\hat{\xi};\hat{\alpha},\hat{\beta}) = \sum\limits_{i=1}^{N} \hat{\alpha}_i y_i = 0$

$\nabla_\xi L(\hat{w},\hat{b},\hat{\xi};\hat{\alpha},\hat{\beta}) = C - \alpha - \beta = 0$

④ $\begin{cases} \hat{\alpha}_i (1 - \hat{\xi}_i - y_i(\hat{w}^T x_i + \hat{b})) = 0 \\ \hat{\beta}_i \hat{\xi}_i = 0 \end{cases}$

then. Dual Form $\Rightarrow$
$$\begin{cases} \max\limits_{\alpha,\beta} \min\limits_{w,b,\xi} L(w,b,\xi;\alpha,\beta) \\ \text{s.t} \quad \alpha \geq 0,\ \beta \geq 0 \end{cases}$$

$$\Rightarrow \begin{cases} \max\limits_{\alpha} & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N}\alpha_i \\ s.t & \alpha \geq 0,\ C-\alpha \geq 0,\ \sum_{i=1}^{N}\alpha_i y_i = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \max\limits_{\alpha} & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N}\alpha_i \\ s.t & 0 \leq \alpha \leq C,\ \sum_{i=1}^{N}\alpha_i y_i = 0 \end{cases}$$

$\Downarrow$

Solve for $\hat{\alpha}$ $\Rightarrow$

① $\hat{\beta} = C - \hat{\alpha}$

② $\hat{w} = \sum_{i=1}^{N}\hat{\alpha}_i y_i x_i$

③ for those $0 < \alpha_s < C$,

we have $y_s(\hat{w}^T x_s + \hat{b}) = 1$ ← from CS condition

④ $\Downarrow$

$$\hat{b} = y_s - \hat{w}^T x_s = y_s - \sum_{i=1}^{N}\hat{\alpha}_i y_i x_i^T x_s$$

prediction

$$\hat{f}(x) = sgn(\hat{w}^T x + \hat{b}) = sgn\left(\sum_{i=1}^{N}\hat{\alpha}_i y_i x_i^T x + \hat{b}\right)$$

where $\hat{b}$ is given by $\hat{b} = y_s - \sum_{i=1}^{N}\hat{\alpha}_i y_i x_i^T x_s$

Another formulation

$$\begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \\ s.t & y_i(w^T x_i + b) \geq 1 - \xi_i \quad i = 1,2,\dots,N \\ & \xi_i \geq 0 \quad i = 1,2,\dots,N \end{cases}$$

Unconstrained Optimization Problem Formulation

$$\Longleftrightarrow \min_{w,b} \frac{1}{2}w^T w + C\sum_{i=1}^{N}\max\{0, 1 - y_i(w^T x_i + b)\}$$

loss function (hinge loss)

Model: $\hat{y} = sgn(\hat{w}^T x + \hat{b})$

$Loss_{hinge}(z) = \max\{0, 1-z\}$

$z = y \cdot (w^T x + b)$ → agreement