Support Vector Machines L11

Hard margin

$$\min_{\underline{w}, w_0} \frac{1}{2} \underline{w}^T \underline{w}$$

s.t. $y_n(\underline{w}^T \underline{x}_n + w_0) \geq 1$

Soft-margin

$$\min_{\underline{w}, w_0, \xi_n} \frac{1}{2} \underline{w}^T \underline{w} + C \sum_n \xi_n$$

s.t. $y_n(\underline{w}^T \underline{x}_n + w_0) \geq 1 - \xi_n$

$\xi_n \geq 0$

Today

① Solve dual form of the max-margin problem, allow discriminant as

$$\boxed{\hat{y} = \begin{cases} +1 & h(\underline{x}, \underline{w}, w_0) > 0 \\ -1 & \text{o.w.} \end{cases}} \quad \left(\sum_n \alpha_n y_n \underline{x}_n^T \underline{x}\right) + w_0$$

with $\alpha_n \geq 0$

② "kernel trick", access basis $\Phi : \mathbb{R}^D \to \mathbb{R}^M$ via $\phi(\underline{x}_n)^T \phi(\underline{x})$

Only use ϕ in this scalar product form.

We have (hard-margin):

(A) $\min\limits_{\underline{w}, w_0} \frac{1}{2}\underline{w}^T\underline{w}$ s.t. $y_n(\underline{w}^T\underline{x}_n + w_0) \geq 1$, all n

Introduce $\alpha_n \geq 0$, each n. Write down in Lagrangian form (for an "≥ 1." form)

(B) $$\min_{\underline{w}, w_0}\left[\max_{\underline{\alpha}\geq 0} \underbrace{\frac{1}{2}\underline{w}^T\underline{w} - \sum_n \alpha_n\left(y_n(\underline{w}^T x + w_0) - 1\right)}_{L(\underline{w}, \underline{\alpha}, w_0)}\right]$$

Lagrangian function

Claim Optimal (A) = Optimal (B)

1) Opt. solution to (B) must satisfy constraints of (A)
2) Opt. solution to (B) sets $\alpha_n = 0$ on all $\underline{x}_n$ with $y_n(\underline{w}^T\underline{x}_n + w_0) > 1$
3) By (1) and (2), opt. solution to (B) satisfies $\alpha_n(y_n(w^T\underline{x}_n + w_0) - 1) = 0$ for all n, and therefore solves (A).

Weak duality

$$\min_{\underline{w},\, w_0} \left[\max_{\underline{\alpha} \geq 0} L(\underline{w}, \underline{\alpha}, w_0) \right] \geq \max_{\underline{\alpha} \geq 0} \left[\min_{\underline{w},\, w_0} L(\underline{w}, \underline{\alpha}, w_0) \right]$$

Strong duality

$$\text{LHS} = \text{RHS}$$

[Duality theory. Holds because objective in (A) is quadratic, constraints are linear]

Dual formulation

$$\max_{\underline{\alpha} \geq 0} \left[\min_{\underline{w},\, w_0} \overbrace{\frac{1}{2} \underline{w}^T \underline{w} - \sum_n \alpha_n \left(y_n (\underline{w}^T \underline{x}_n + w_0) - 1 \right)}^{OBJ} \right]$$

Can now write in terms of $\underline{\alpha}$ only

For any $\underline{\alpha}$, optimal $\underline{w}$, w_0 must satisfy:

$$\frac{\partial L(\underline{w}, \underline{\alpha}, w_0)}{\partial \underline{w}} = \underline{w} - \sum_n \alpha_n y_n \underline{x}_n = 0$$

$$\Leftrightarrow \quad \underline{w} = \sum_n \alpha_n y_n \underline{x}_n \qquad (*)$$

$\min_{w_0} - w_0 \sum \alpha y$

$- w_0 (-1)$

$+ w_0$

For any $\underline{\alpha}$, optimal $\underline{w}, w_0$ ~~st~~ satisfies:

$$\frac{\partial L(\underline{w}, \underline{\alpha}, w_0)}{\partial w_0} = -\sum_n \alpha_n y_n = 0 \qquad (\square)$$

Note: If $\sum \alpha_n y_n < 0$, then $w_0 \to -\infty$ makes $\min_{w, w_0} [\cdot]$ arbitrarily small

Simplify OBJ

$$\frac{1}{2} \underline{w}^T \underline{w} - \underline{w}^T \sum_n \alpha_n y_n \underline{x}_n - \cancel{w_0 \sum_n \alpha_n y_n}^{\,0} + \sum_n \alpha_n$$

$$= -\frac{1}{2} \underline{w}^T \underline{w} + \sum_n \alpha_n \quad \{\text{subst. } (*), \text{ adding constraint } (\square)\}$$

$$= -\frac{1}{2} \left(\sum_n \alpha_n y_n \underline{x}_n \right)^T \left(\sum_{n'} \alpha_{n'} y_{n'} \underline{x}_{n'} \right) + \sum \alpha_n$$

Note: If $\sum \alpha_n y_n > 0$, then $w_0 \to +\infty$ makes $\min_{w, w_0} [\cdot]$ arbitrarily small

Dual, Hard-margin formulation

$$\max_{\underline{\alpha} \geq 0} \sum_n \alpha_n - \frac{1}{2} \sum_n \sum_{n'} \alpha_n \alpha_{n'} y_n y_{n'} \underline{x}_n^T \underline{x}_{n'}$$

$$\text{s.t.} \sum_n \alpha_n y_n = 0, \quad \alpha_n \geq 0$$

Soft-margin: Same as above, except place an upper-bound on α_n: $C \geq \alpha_n \geq 0$

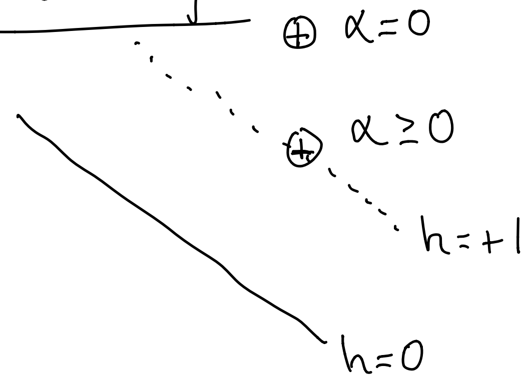
(Prevents dual becoming unbounded)

Notes

Discriminant $h(\underline{x}, \underline{\alpha}, w_0) = \sum_n \alpha_n y_n \underline{x}_n^T \underline{x} + w_0$

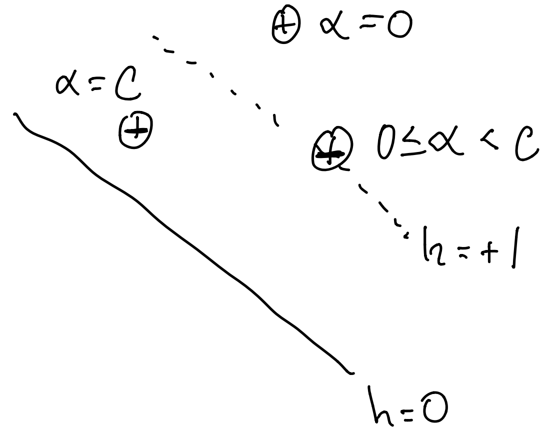
Support vectors: $Q = \{ \alpha_n : \alpha_n > 0 \}$

Hard-margin



$\alpha_n > 0 \Rightarrow$ example on the margin boundary

Soft margin



$0 < \alpha_n < C$

$\Rightarrow$ example on margin boundary

Solve for w_0

Recall $y_n(\underline{w}^T \underline{x}_n + w_0) = 1$ for any $\underline{x}_n$ on margin boundary.

Find any $\underline{x}_n$ on boundary, solve for w_0

Why is dual formulation useful?

Consider basis function $\phi : \mathbb{R}^D \to \mathbb{R}^M$

$$\max_{\underline{\alpha}} \sum_n \alpha_n - \frac{1}{2} \sum_n \sum_{n'} \alpha_n \alpha_{n'} y_n y_{n'} \overbrace{\phi(\underline{x}_n)^T \phi(\underline{x}_{n'})}^{K(\underline{x}_n, \underline{x}_{n'})}$$

$$\text{s.t. } \sum_n \alpha_n y_n = 0, \quad C \geq \alpha_n \geq 0 \quad \text{all } n$$

"kernelized" form

Similarly, the discriminant:

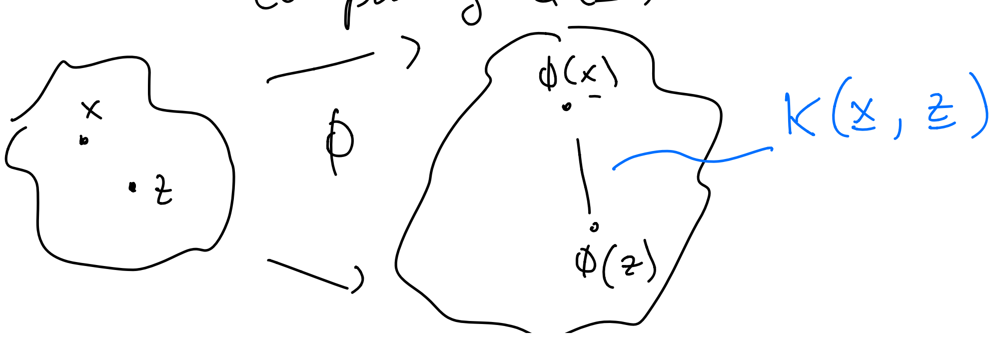
$$h(\underline{x}, \underline{\alpha}, w_0) = \sum_n \alpha_n y_n \underbrace{\phi(\underline{x}_n)^T \phi(\underline{x})}_{K(\underline{x}_n, \underline{x})}$$

Only need kernel function

$$K(\underline{x}, \underline{z}) = \phi(\underline{x})^T \phi(\underline{z})$$

"Kernel trick"

Compute $K(\underline{x}, \underline{z})$ without computing $\phi(\underline{x})$ or $\phi(\underline{z})$



Example Quadratic kernel

$$K_{quad}(\underline{x}, \underline{z}) = (\underline{x}^T \underline{z})^2$$

Suppose $\underline{x}, \underline{z} \in \mathbb{R}^2$, $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $\underline{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$

$$K_{quad}(\underline{x}, \underline{z}) = (\underline{x}^T \underline{z})^2 = (x_1 z_1 + x_2 z_2)^2$$

$$= x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2$$

$$= \begin{bmatrix} x_1^2 & x_1 x_2 & x_2 x_1 & x_2^2 \end{bmatrix} \begin{bmatrix} z_1^2 \\ z_1 z_2 \\ z_2 z_1 \\ z_2^2 \end{bmatrix} = \phi(\underline{x})^T \phi(\underline{z})$$

Corresponds to a basis that uses all degree-2 terms

$$K(\underline{z}, \underline{z}) = (1 + \underline{x}^T \underline{z})^2$$

correspond to also including linear + constant term

Example Polynomial kernel

$$K_{poly}(\underline{x}, \underline{z}) = (1 + \underline{x}^T \underline{z})^q, \text{ integer } q \geq 2$$

↳ including all terms up to degree q

↳ as of the basis has $O(D^q)$ terms

<u>Example</u> Gaussian kernel

$$K_{Gauss}(\underline{x}, \underline{z}) = \exp\left[-\frac{\|\underline{x}-\underline{z}\|_2^2}{\lambda}\right],$$

bandwidth $\lambda > 0$

decays exponentially in squared distance

↳ Corresponds to an ∞-dimensional basis!

Crucial idea:

Don't expand $\underline{x}$ to $\phi(\underline{x})$ + take inner products. Just do direct calculation

Notes

1. Kernel engineering

(What is a valid kernel function k?)

Function K defines a kernel (or "Gram" matrix) $\underline{K}$ on data $\{\underline{x}_1, \ldots, \underline{x}_N\}$

$$N \overset{N}{\begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}} \quad \underline{K}_{n,n'} = K(\underline{x}_n, \underline{x}_{n'})$$

Mercer's Theorem

Kernel function k is valid if & only if Gram matrix is p.s.d.

Allows engineering

Suppose valid K_1, valid K_2,

then all the following are valid:

aK_1, $a > 0$

$K_1 + K_2$

$\text{poly}(K(\cdot,\cdot))$

$\exp(K(\cdot,\cdot))$

$f(\underline{x})K(\underline{x},\underline{z})f(\underline{z})$ any function f

$\vdots$