SVM

Given $(\underline{x}, y) \sim \mathcal{D}$. $\underline{x} \in \mathcal{X}$, $y \in \mathcal{Y}$

Task: Construct a prediction rule $f : \mathcal{X} \to \mathcal{Y}$

Classification : $\mathcal{Y} = \{-1, 1\}$

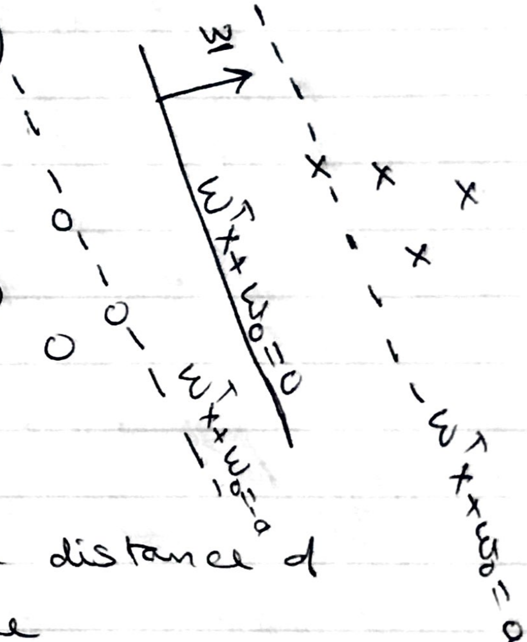
Regression as classification

$$\hat{y} = \text{sign}(\underbrace{f(\underline{x})}_{\text{score of } \underline{x} \text{ under } f(\cdot)})$$

Linear classifiers :

$$f(x) = \underline{w}^T x + w_0$$ parametrized by $d+1$ variables.

$$\hat{y} = \text{sign}(w^T x + w_0)$$



Basically assign label based on which side of the hyperplane ($\underline{w}^T \underline{x} + w_0 = 0$) the point $x \in \mathbb{R}^d$ lies.

Score $(\underline{x}) = f(x) = w^T x + w_0$ is proportional to the distance of point x from hyperplane

Score establishes a 1-D coordinate system by projecting onto vector $\underline{w}$.

Learning : Given $\{(x_i, y_i)\}_{i=1}^n$ how do we find good w.

(1) Least squares regression (with regularization)

$$\hat{w} = \arg\min \sum_{i=1}^n (y_i - w^T x_i)^2 + \lambda \|w\|^2$$

Sensitive to imbalance

(2) Logistic regression (with regularization)

$$\hat{w} = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} -\log p(y_i | x_i, w) + \frac{1}{2\sigma^2} \|w\|^2$$

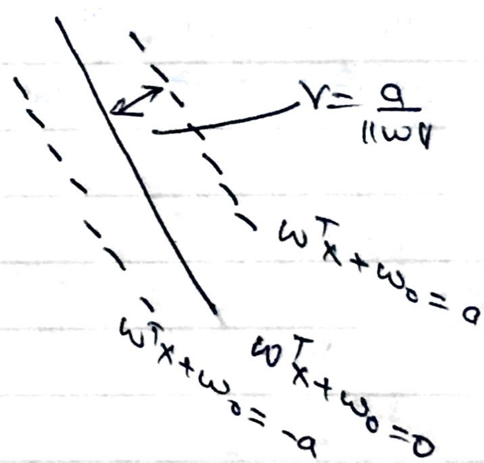
$$\text{where } p(y_i | x_i ; w) = \begin{cases} \sigma(w^T x_i + w_0), & \text{if } y_i = 1 \\ 1 - \sigma(w^T x_i + w_0), & \text{if } y_i = 0 \end{cases}$$

$$= \sigma(w^T x_i + w_0)^{y_i} \left(1 - \sigma(w^T x_i + w_0)\right)^{1 - y_i}$$

(3) SVM (max margin $\Rightarrow$ regularization)

$$\max_{\underline{w}} \frac{a}{\|\underline{w}\|}$$

s.t. $y_i(\underline{w}^T \underline{x}_i + w_0) \geq a$
$\forall i = 1, \dots, n$



Scale $\underline{w} \mapsto a\underline{w}$
$w_0 \mapsto a w_0$
$\Rightarrow$ Objective invariant

$$\Rightarrow \max_{\underline{w}} \frac{1}{\|\underline{w}\|}$$

s.t. $y_i(\underline{w}^T \underline{x}_i + w_0) \geq 1 \quad \forall i = 1, \dots, n$

Equivalently

$$\Rightarrow \quad \min_{\underline{w}} \frac{1}{2}\|\underline{w}\|^2 \quad \text{s.t.} \quad y_i(\underline{w}^T \underline{x}_i + w_0) \geq 1 \quad \forall i = 1, \dots, N$$

Loss for each constraint violation

$$\max_{\alpha_i \geq 0} \alpha_i \left(1 - y_i (w^T x_i + w_0)\right)$$

If constraint met loss $= 0$
otherwise loss $= \infty$

Makes sense when data linearly separable

$$\Rightarrow \quad \min_{\underline{w}} \left\{ \frac{1}{2}\|\underline{w}\|^2 + \sum_{i=1}^{N} \max_{\alpha_i \geq 0} \alpha_i \left[1 - y_i(\underline{w}^T \underline{x}_i + w_0)\right] \right\}$$

$$\Rightarrow \quad \min_{\underline{w}} \max_{\underline{\alpha} \geq 0} \left\{ \frac{1}{2}\|\underline{w}\|^2 + \sum_{i=1}^{n} \alpha_i \left[1 - y_i(\underline{w}^T \underline{x}_i + w_0)\right] \right\}$$

$$\Rightarrow \quad \max_{\underline{\alpha} \geq 0} \min_{\underline{w}} \left\{ \frac{1}{2}\|\underline{w}\|^2 + \sum_{i=1}^{n} \alpha_i \left[1 - y_i(\underline{w}^T \underline{x}_i + w_0)\right] \right\}$$

$\uparrow$

Duality under KKT conditions ($d^* = p^*$)

SVM

$$\max_{\alpha \geq 0} \min_{\underline{w} \in \mathbb{R}^d} \underbrace{\frac{1}{2}\|\underline{w}\|^2 + \sum_{i=1}^{n} \alpha_i \left[1 - y_i(\underline{w}^T \underline{x}_i + w_0)\right]}_{\mathcal{L}(\underline{w}, \alpha_1, \ldots, \alpha_n)}$$

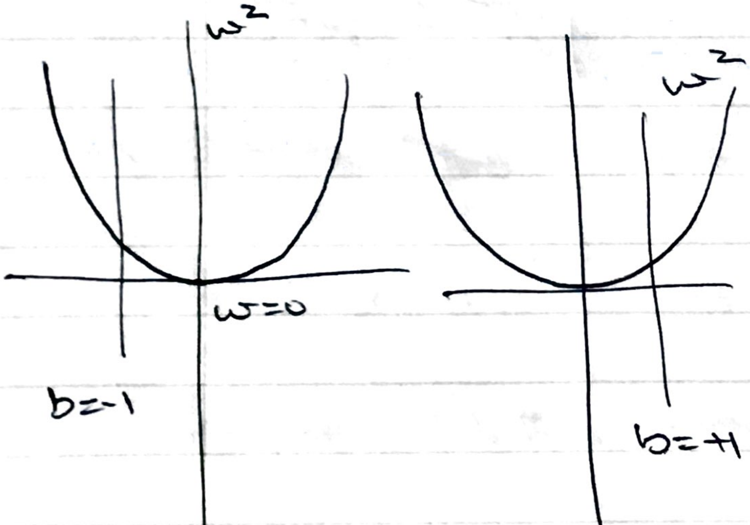
- Fix $\alpha \geq 0$
- Minimize over $\underline{w}$ to get $\underline{w}(\alpha)$, $w_0(\alpha)$
- Maximize over $\alpha \geq 0$ the function $J(\underline{w}(\alpha), w_0(\alpha); \alpha)$ to get $\alpha^* \geq 0$.

- Find $\underline{w}(\alpha^*)$, $w_0(\alpha^*)$ by substitution.

Eg:

Primal

$$\min_{w \in \mathbb{R}} w^2 \quad \text{s.t. } w \geq b$$



Dual

$$\max_{\alpha \geq 0} \min_{w \in \mathbb{R}} \underbrace{w^2 - \alpha(w-b)}_{L(w,\alpha)}$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \cancel{w(\alpha) = \frac{\alpha}{2}}$$

$$\boxed{w(\alpha) = \frac{\alpha}{2}}$$

$$\therefore L(w(\alpha), \alpha) = -\frac{\alpha^2}{4} + b\alpha$$

$$\frac{\partial L(w(\alpha), \alpha)}{\partial \alpha} = -\frac{\alpha}{2} + b$$

$$\boxed{\alpha^* = \max(0, 2b)}$$

$$\boxed{w^* = w(\alpha^*) = \max(0, b)}$$

Back to SVMs

$$\underline{w}(\underline{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i \underline{x}_i$$

Plugging back to the Lagrangian get QP

$$\max_{\substack{\alpha \geq 0 \\ \sum \alpha_i y_i = 0}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \right\}$$

Solve it to get α^*, plug back in to get

$$\underline{w}^* = \underline{w}(\alpha^*) = \sum_{i=1}^{n} \alpha_i^* y_i \underline{x}_i$$

Support vectors

$$\alpha_i^* > 0 \quad \text{iff} \quad y_i(w \cdot x_i + w_0) = 1$$

$$\alpha_i^* = 0 \quad \text{if} \quad y_i(w \cdot x_i + w_0) > 1$$

SVM classifier

$$\underline{w}^* = \sum_{\alpha_i^* > 0} \alpha_i^* y_i \underline{x}_i$$

Non-separable case (linearly)

We can no ~~long~~ longer satisfy

$$y_i(\underline{w} \cdot x_i + w_0) \geq 1 \quad \forall i$$

which means the objective $= \infty$.

Modify the loss, ~~define~~ consider

$$\xi_i = \max\{0, 1 - y_i(w \cdot x_i + w_0)\}$$

$$\max_{C \geq \alpha_i \geq 0} \alpha_i \left(1 - y_i(\underline{w} \cdot \underline{x}_i + w_0)\right) \quad \begin{cases} 0 \text{ if constraint met} \\ C \cdot \text{"violation" otherwise} \end{cases}$$

$$\equiv C \underbrace{\max\{0, 1 - y_i(w \cdot x_i + w_0)\}}_{\xi_i \geq 0}$$

- Primal: $\min_{w} \left\{ \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \right\}$

  s.t $\xi_i \geq 0$

Dual $\max \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \right\}$

s.t. $\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \quad \forall i = 1, \ldots, n$