Today: SVMs continued

Hard-margin SVM:

$$\text{minimize} \quad \frac{1}{2}\|\omega\|_2^2$$

$$\text{s.t.} \quad y_i(\vec{\omega}^T\vec{x}_i + b) \geq 1 \quad \forall i$$

Soft-margin SVM:

$$\text{minimize} \quad \frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^{n}\xi_i$$

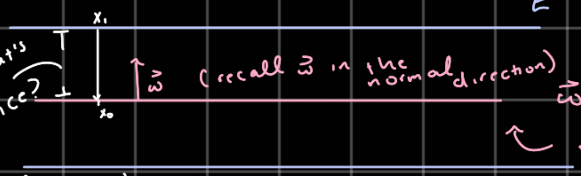
$$\text{s.t.} \quad y_i(\vec{\omega}^T\vec{x}_i + b) \geq 1 - \xi_i \quad \forall i \quad] \text{ Dual variable } \alpha_i$$

$$\xi_i \geq 0 \quad] \text{ Dual variable } \beta_i$$

$(1-\xi_i) - y_i(\vec{\omega}^T\vec{x}_i + b) \leq 0$

$-\xi_i \leq 0$

Classifier: $\Rightarrow$ Label $= \begin{cases} +1 & \omega^T x + b > 0 \\ -1 & o/w \end{cases}$



Q/What's this distance?

↳ Travel in direction of $\vec{\omega}$ to get the orthogonal projection

↳ Q/ How far across $\vec{\omega}$ to get to the next level?

↳ if $\vec{x}_0 \in \vec{\omega}^T\vec{x} + b = 0$

Let $\vec{x}_1 = \vec{x}_0 + k\vec{\omega}$ s.t. $\vec{x}_1 \in \vec{\omega}^T\vec{x} + b = 1$

$$\vec{\omega}^T\vec{x}_1 + b = 1 \Rightarrow \vec{\omega}^T(\vec{x}_0 + k\vec{\omega}) = 1 - b$$

$$\underbrace{\vec{\omega}^T\vec{x}_0}_{-b} + k\vec{\omega}^T\vec{\omega} = 1 - b$$

$$-b + k\|\vec{\omega}\|_2^2 = 1 - b$$

$$k\|\vec{\omega}\|_2^2 = 1$$

$$k = \frac{1}{\|\omega\|_2}$$

$$\|x_0 - x_1\|_2 = k\|\vec{\omega}\|_2 = \frac{1}{\|\omega\|_2}$$

Dual perspective

$$\mathcal{L}(\vec{\omega}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}\|\vec{\omega}\|_2^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\left(\alpha_i(1-\xi_i) - y_i(\omega^T x_i + b)\right) + \sum_{i=1}^{n}\beta_i(-\xi_i)$$

$$\mathcal{L} = \frac{1}{2}\|\omega\|_2^2 - \sum_{i=1}^{n} \alpha_i y_i (\vec{\omega}^T \vec{x}_i + b) + \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} (C - \alpha_i - \beta_i)\xi_i$$

↖ grouping together ξ_i's

$$p^* = \min_{\vec{\omega}, b, \vec{\xi}} \max_{\vec{\alpha}, \vec{\beta}} \mathcal{L}$$

$$d^* = \max_{(\text{dual})} \min_{(\text{primal})} \mathcal{L}$$

↳ strong duality holds → KKT conditions necessary & sufficient

↙ use

First order KKT conditions

$$\nabla_{\vec{\omega}} \mathcal{L} = \vec{\omega} - \sum \alpha_i y_i \vec{x}_i = 0$$

↳ $\vec{\omega} = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i$

dual variable

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

(+1) & (-1) data points

"optimal sum of weighted dual variables equals 0"

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \rightarrow \beta_i = C - \alpha_i$$

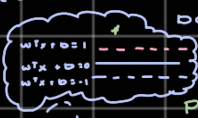
Note: this is because, in our original prob, we're minimizing

$$\frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^{n} \xi_i$$

s.t.

$$y_i(\vec{\omega}^T \vec{x}_i + b) \geq 1 - \xi_i$$

Since we have a minimization, we know that the level set associated with being on the boundary is $y_i \omega^T x_i + b = 1$. Since we know that any point past the boundary is also positive (/labeled as +1) we have no incentive to mak ξ_i nonzero, because we're trying to minimize



FURTHER EXPLANATION

Complementary Slackness

$$\alpha_i(1 - \xi_i - y_i(\omega^T x_i + b)) = 0 \qquad \forall i$$

$$\beta_i(\xi_i) = 0$$

↳ examine dual variables:

$$\alpha_i(1 - \xi_i - y_i(\omega^T x_i + b)) = 0 \qquad \forall i$$

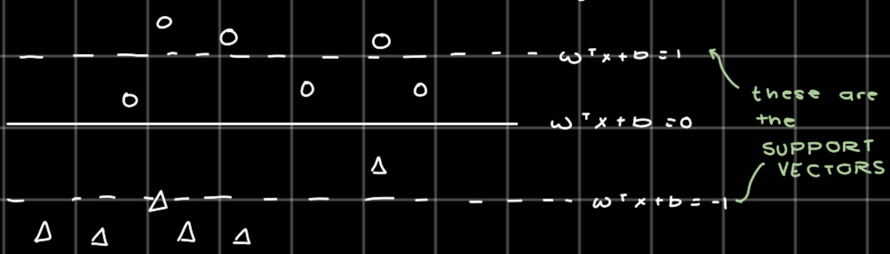
Case 1: $\alpha_i \neq 0$, $0 < \alpha_i < C$

$$(1 - \xi_i) - y_i(\omega^T x_i + b) = 0$$

$$\alpha_i \neq C \Rightarrow \beta_i = C - \alpha_i \Rightarrow \beta_i \neq 0 \Rightarrow \beta_i \xi_i = 0 \Rightarrow \xi_i = 0$$

This implies that the points are on the margin or far away from it (i.e, not violating the margin)

$\Rightarrow$ $1 = y_i(\vec{\omega}^T \vec{x}_i + b)$ $\Rightarrow$ all points are exactly on the margin



case 2: $\alpha_i \neq 0$, $\alpha_i = c$

$$\cancel{c} - \cancel{\alpha_i} - \beta_i = 0 \implies \beta_i = 0$$

↳ if $\beta_i = 0$ the ξ_i need not be 0.

$$\implies (1 - \xi_i) - y_i(w^T x + b) = 0 \implies y_i(w^T x_i + b) = 1 - \xi_i \leq 1$$

case 3: $\alpha_i = 0$

$$\alpha_i = 0 \implies c - \alpha_i - \beta_i = 0 \implies c = \beta_i \implies \xi_i = 0$$

$\implies$ no margin violation

Let's compute the dual:

$$\mathcal{L}^* = \frac{1}{2}\|w\|_2^2 - \underbrace{\sum_{i=1}^n \alpha_i y_i(\vec{w}^T\vec{x}_i + b)}_{= \sum_{i=1}^n \alpha_i y_i \vec{w}^T\vec{x}_i + b\cancel{\sum_{i=1}^n \alpha_i y_i}^{0}} + \sum_{i=1}^n \alpha_i + \underbrace{\sum_{i=1}^n \cancel{(c - \alpha_i - \beta_i)}\xi_i}_{\text{By FOC, at optimality, this term equals 0}}$$

$$= \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^n \alpha_i y_i \vec{w}^T\vec{x}_i + \sum_{i=1}^n \alpha_i$$

$$= \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^n \alpha_i (y_i w^T x_i - 1)$$

$$= \frac{1}{2}\|w\|_2^2 - \underbrace{\left(\sum_{i=1}^n \alpha_i y_i \vec{x}_i^T\right)}_{= \vec{w}^T \text{ (FOC)}} w + \sum_{i=1}^n \alpha_i$$

$$= \frac{1}{2}\|w\|_2^2 - w^T w + \sum_{i=1}^n \alpha_i$$

$$= -\frac{1}{2}\|w\|_2^2 + \sum_{i=1}^n \alpha_i$$

$$= -\frac{1}{2}\left(\sum_{i=1}^n \alpha_i y_i \vec{x}_i^T\right)\left(\sum_{i=1}^n (\alpha_i y_i \vec{x}_i^T)\right) + \sum_{i=1}^n \alpha_i$$

$$d^* = \max_{\vec{\alpha}, \beta} \min_{\vec{w}, \beta, \vec{\xi}} \mathcal{L} = \max_{\alpha, \beta} \mathcal{L}^*$$

since we sub in the optimal values with the $\mathcal{L}^*$ $\wedge \doteq$ the min of a constant is a constant, this minimization drops out of the computation

$$\mathcal{L}^* = -\frac{1}{2}[\alpha_1 y_1 \;\cdots\; \alpha_n y_n] \begin{bmatrix} \text{—} & x_1^T & \text{—} \\ & \vdots & \\ \text{—} & x_n^T & \text{—} \end{bmatrix} \begin{bmatrix} | & & | \\ \vec{x}_1 & \cdots & \vec{x}_n \\ | & & | \end{bmatrix} \begin{bmatrix} \alpha_1 y_1 \\ \vdots \\ \alpha_n y_n \end{bmatrix} + \sum \alpha_i$$

$$= -\frac{1}{2}\vec{\alpha}^T \underbrace{\operatorname{diag}(\vec{y}) \cdot XX^T \cdot \operatorname{diag}(\vec{y})}_{\doteq Q} \vec{\alpha} + \sum \alpha_i$$

if we're using a kernel, this would be replaced by $K(XX^T)$ → easy to compute pairwise inner product

$$d^* = \max_{\vec{\alpha}, \vec{\beta}} -\frac{1}{2}\vec{\alpha}^T Q \vec{\alpha} + \sum \alpha_i$$

s.t. $\sum \alpha_i y_i = 0$

$\left.\begin{array}{l} c - \alpha_i - \beta_i = 0 \quad \forall i \\ \alpha_i \geq 0 \\ \beta_i \geq 0 \end{array}\right\} 0 \leq \alpha_i \leq c$

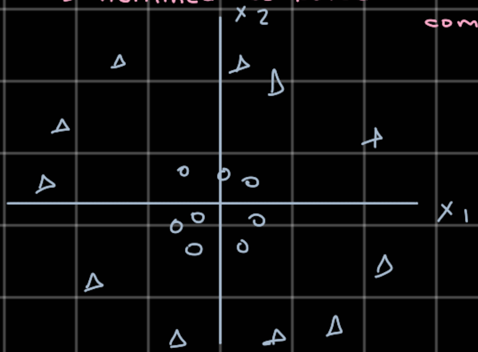
KERNEL

$F(\vec{x}) = \vec{\omega}^T \vec{x} + b$ if $F(\vec{x}_i) > 0 \Rightarrow y_i = +1$

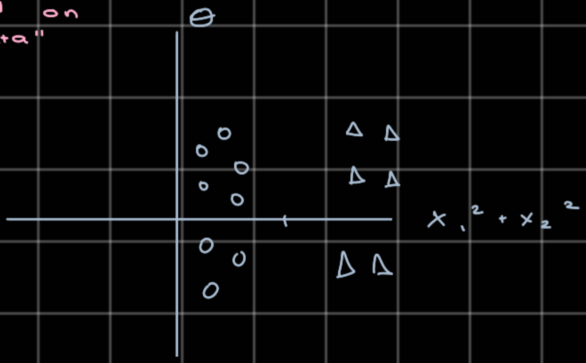
$f(\vec{x}) = \vec{\omega}\, \underbrace{\phi(\vec{x})}_{} + b$

some featurization of $\vec{x}$

↳ "nonlinear computed feature based on computed data"



⇒



eg: $\Phi(x) = e^{-\gamma(x_1^2 + x_2^2)}$