maximize $\quad 1^T.\alpha - \frac{1}{2}\alpha^T.(K \circ (yy^T))\alpha$

subject to: $\quad y^T.\alpha = 0$

$C.1 \geq \alpha \geq 0$

in order to have a concave objective function

K should be a psd (positive semidefinite) matrix

$\underbrace{a^T}_{1\times N} . \underbrace{K}_{N\times N} \underbrace{a}_{N\times 1} \geq 0 \qquad \forall a \in \mathbb{R}^N$ $\Rightarrow$ all eigenvalues of K should be nonnegative.

Constructing kernels:

$k(x_i, x_j) \Rightarrow c.k(x_i, x_j)$

$\Downarrow$

$c > 0$

$a^T.Ka \geq 0 \Rightarrow a^T(cK)a = \underbrace{c}_{>0} . \underbrace{a^T.Ka}_{\geq 0} \geq 0$

$k_1(x_i, x_j)$

$k_2(x_i, x_j)$ $\Rightarrow k_1(x_i, x_j) + k_2(x_i, x_j)$

$\left.\begin{array}{l} a^T K_1 a \geq 0 \\ a^T K_2 a \geq 0 \end{array}\right\} \quad a^T(K_1 + K_2)a \geq 0$

$\underbrace{a^T K_1 a}_{\geq 0} + \underbrace{a^T K_2 a}_{\geq 0}$

Exercise: Is $k_1(x_i, x_j)k_2(x_i, x_j)$ a valid kernel?

Multiclass Kernel Machines

$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N} \qquad y_i \in \{1, 2, \ldots, K\}$

one-versus-all (OVA) $\Rightarrow$

1^{+} vs $\{2, 3, \ldots, K\}^{-} \Rightarrow \text{SVM}_1$

2^{+} vs $\{1, 3, \ldots, K\}^{-} \Rightarrow \text{SVM}_2$

3^{+} vs $\{1, 2, 4, \ldots, K\}^{-} \Rightarrow \text{SVM}_3$

$\vdots \quad \vdots \qquad \vdots \qquad \vdots$

K^{+} vs $\{1, 2, \ldots, K-1\}^{-} \Rightarrow \text{SVM}_K$

\# of classifiers $= K$

training set size $= N$

test data point x^*

$f_1(x^*)$ (SVM$_1$) $\quad f_2(x^*)$ (SVM$_2$) $\quad - - - - \quad f_K(x^*)$ (SVM$_K$) $\Rightarrow$ pick the maximum one

one-versus-other (OVO) $\Rightarrow$

1^{+} vs $2^{-} \Rightarrow \text{SVM}_{1\text{vs}2}$

1^{+} vs $3^{-} \Rightarrow \text{SVM}_{1\text{vs}3}$

$\vdots$

$(K-1)^{+}$ vs $K^{-} \Rightarrow \text{SVM}_{(K-1)\text{vs}K}$

\# of classifiers $= \frac{K(K-1)}{2}$

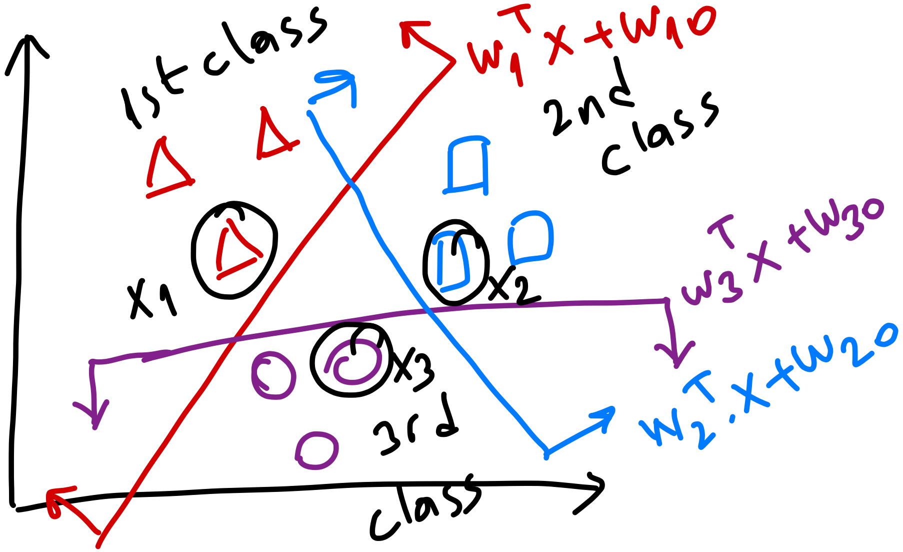
training set size $= 2 \cdot \frac{N}{K}$ (assuming classes are equal-sized)

x^*

$f_{1\text{vs}2}(x^*) \quad f_{1\text{vs}3}(x^*) \quad \ldots \quad f_{(K-1)\text{vs}K}(x^*)$

$\Downarrow$

pick the one with the maximum # of wins.



$$w_1^T . x_1 + w_{10} \geq w_2^T x_1 + w_{20} + 2 - \varepsilon_{12}$$

$$w_1^T . x_1 + w_{10} \geq w_3^T . x_1 + w_{30} + 2 - \varepsilon_{13}$$

$$w_2^T . x_2 + w_{20} \geq w_1^T . x_2 + w_{10} + 2 - \varepsilon_{21}$$

$$w_2^T x_2 + w_{20} \geq w_3^T . x_2 + w_{30} + 2 - \varepsilon_{23}$$

minimize $\frac{1}{2} \sum_{c=1}^{K} ||w_c||_2^2 + C \sum_{i=1}^{N} \sum_{c=1}^{K} \varepsilon_{ic}$

subject to: $w_{y_i}^T x_i + w_{y_i 0} \geq w_c^T x_i + w_{c0} + 2 - \varepsilon_{ic}$ $\forall (i, c \neq y_i)$

$\varepsilon_{ic} \geq 0$ $\forall (i, c \neq y_i)$

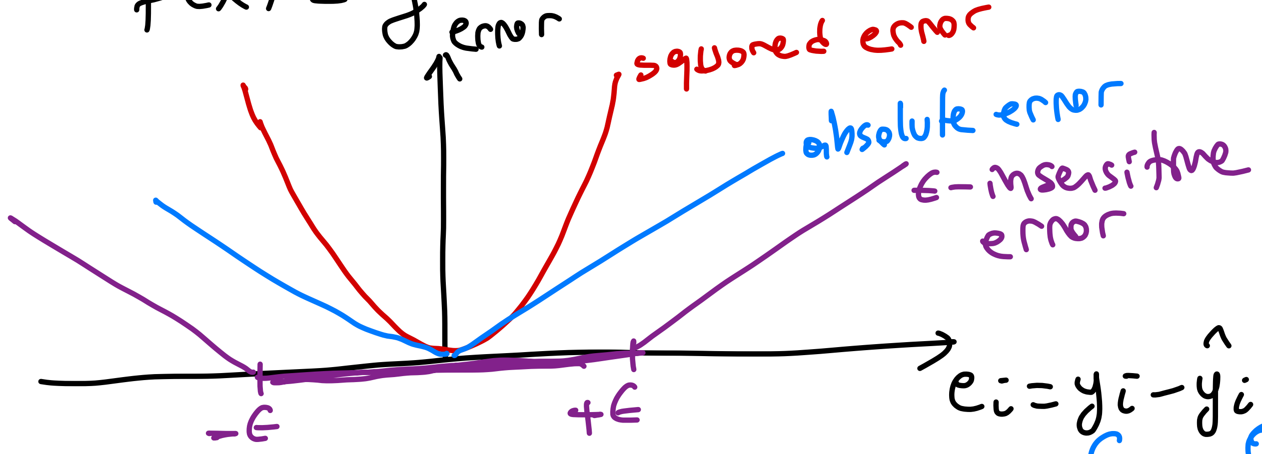
# of decision variables $= (D+1)K + N(K-1)$

# of constraints $= N(K-1)$

Kernel machines for Regression

$f(x) = \hat{y} = w^T . x + w_0$

squared error $\Rightarrow \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$



$$\epsilon\text{-insensitive loss} = \begin{cases} 0 & \text{if } |y_i - \hat{y}_i| \leq \epsilon \\ |y_i - \hat{y}_i| - \epsilon & \text{otherwise} \end{cases}$$

minimize $\quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{N} (\xi_i^+ + \xi_i^-)$

subject to:

$$y_i - [w^T . x_i + w_0] \leq \epsilon + \xi_i^+ \quad \forall i \quad] \ \alpha_i^+ \checkmark$$

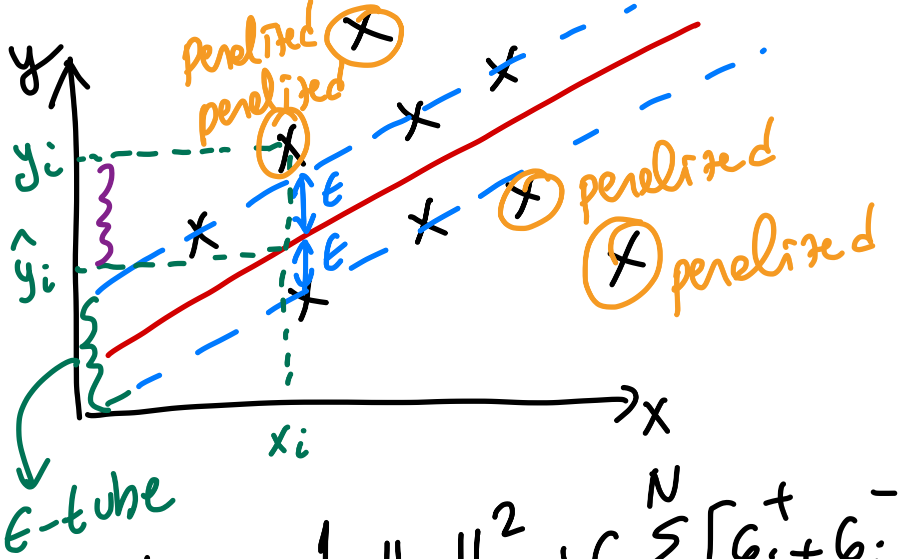
$$[w^T . x_i + w_0] - y_i \leq \epsilon + \xi_i^- \quad \forall i \quad] \ \alpha_i^- \checkmark$$

$$\xi_i^+ \geq 0 \quad \forall i \quad] \ \beta_i^+$$

$$\xi_i^- \geq 0 \quad \forall i \quad] \ \beta_i^-$$

$\#$ of decision variables $= \underbrace{D}_{w} + \underbrace{1}_{w_0} + \underbrace{2N}_{\xi_i^+, \xi_i^-}$

$\#$ of constraints $= 2 . N$



$$L_p = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N}\left[\xi_i^+ + \xi_i^-\right] - \sum_{i=1}^{N}\alpha_i^+\left[-y_i + w^T x_i + w_0 + \epsilon + \xi_i^+\right]$$
$$- \sum_{i=1}^{N}\alpha_i^-\left[y_i - w^T x_i - w_0 + \epsilon + \xi_i^-\right] - \sum_{i=1}^{N}\beta_i^+\xi_i^+ - \sum_{i=1}^{N}\beta_i^-\xi_i^-$$

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{N}\alpha_i^+ x_i + \sum_{i=1}^{N}\alpha_i^- x_i = 0 \Rightarrow \boxed{w = \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)x_i}$$

$$\frac{\partial L_p}{\partial w_0} = -\sum_{i=1}^{N}\alpha_i^+ + \sum_{i=1}^{N}\alpha_i^- = 0 \Rightarrow \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-) = 0$$

$$\frac{\partial L_p}{\partial \xi_i^+} = C - \alpha_i^+ - \beta_i^+ = 0 \Rightarrow \alpha_i^+ + \beta_i^+ = C \Rightarrow 0 \le \alpha_i^+ \le C$$

$$\frac{\partial L_p}{\partial \xi_i^-} = C - \alpha_i^- - \beta_i^- = 0 \Rightarrow \alpha_i^- + \beta_i^- = C \Rightarrow 0 \le \alpha_i^- \le C$$

$$\text{maximize} \quad \sum_{i=1}^{N} y_i [\alpha_i^+ - \alpha_i^-] - \epsilon \sum_{i=1}^{N} (\alpha_i^+ + \alpha_i^-)$$

$$- \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) \boxed{x_i^T . x_j}$$

$$\text{subject to:} \quad \sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-) = 0 \; ✓$$

$$C \geq \alpha_i^+ \geq 0 \quad \forall i$$

$$C \geq \alpha_i^- \geq 0 \quad \forall i$$

$\#$ of decision variables $= 2N$

$\#$ of constraints $= 1$

replace $x_i^T x_j$ with

$$\Phi(x_i)^T \Phi(x_j) = k(x_i, x_j)$$

new data point

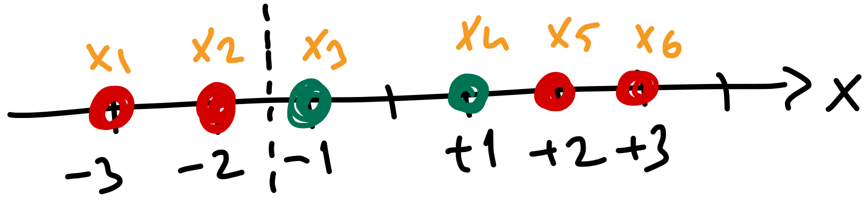
$\downarrow$

x^*

$$\Rightarrow f(x^*) = w^T . x^* + w_0 = \sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-) \underbrace{\underbrace{x_i^T . x^*}_{k(x_i, x)}}_{\Phi(x_i)^T \Phi(x)} + w_0$$

$$x_i \rightarrow \Phi(x_i) = \begin{bmatrix} x_i \\ x_i^2 \end{bmatrix} = z_i$$

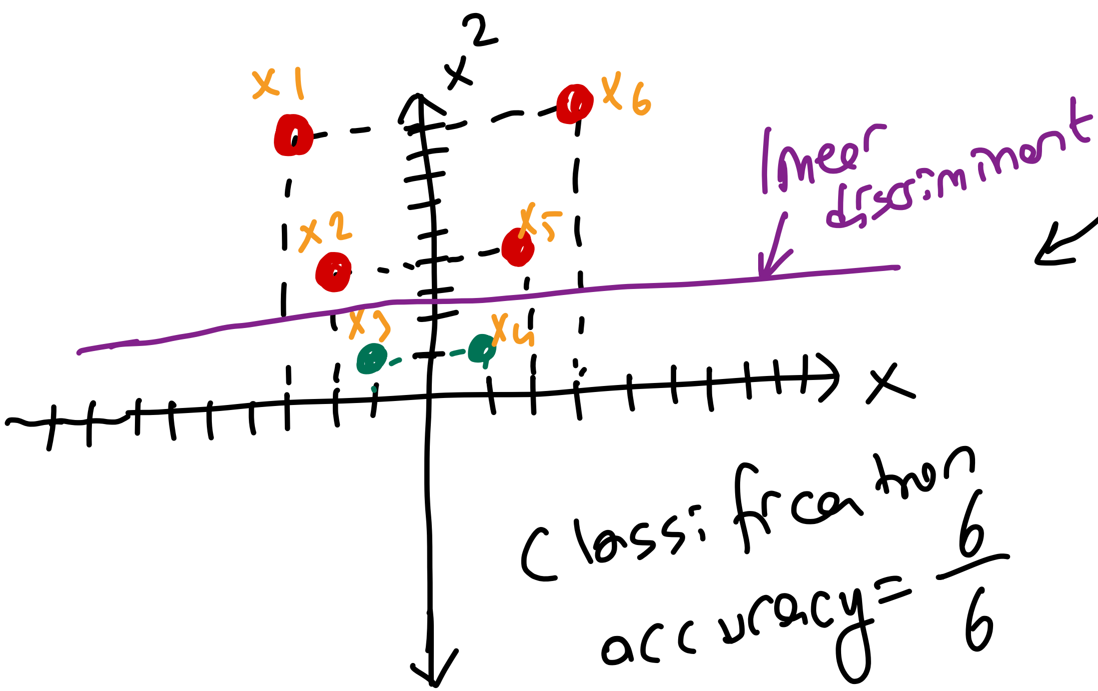
$D = 1 \qquad d = 2$



RED ← | → GREEN

Classification accuracy $= \frac{4}{6}$

Φ



Classification accuracy $= \frac{6}{6}$

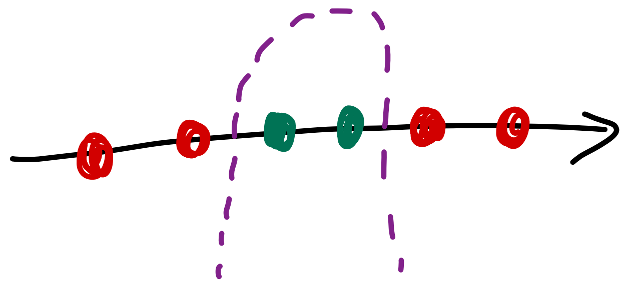
$$-3 \Rightarrow \begin{bmatrix} -3 \\ +9 \end{bmatrix} \qquad +2 \Rightarrow \begin{bmatrix} +2 \\ +4 \end{bmatrix}$$

$$-2 \Rightarrow \begin{bmatrix} -2 \\ +4 \end{bmatrix} \qquad +3 \Rightarrow \begin{bmatrix} +3 \\ +9 \end{bmatrix}$$

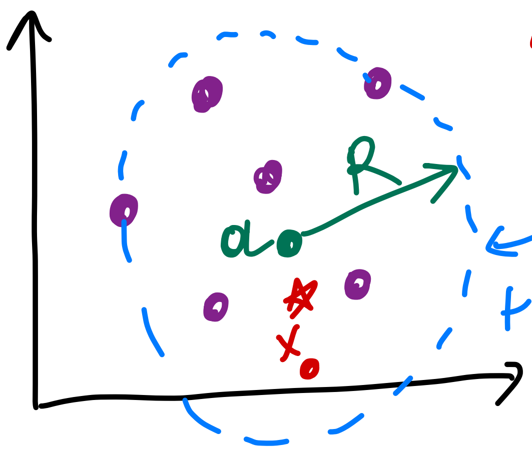
$$-1 \Rightarrow \begin{bmatrix} -1 \\ +1 \end{bmatrix}$$

$$+1 \Rightarrow \begin{bmatrix} +1 \\ +1 \end{bmatrix}$$

Φ^{-1}



One-Class Kernel machines

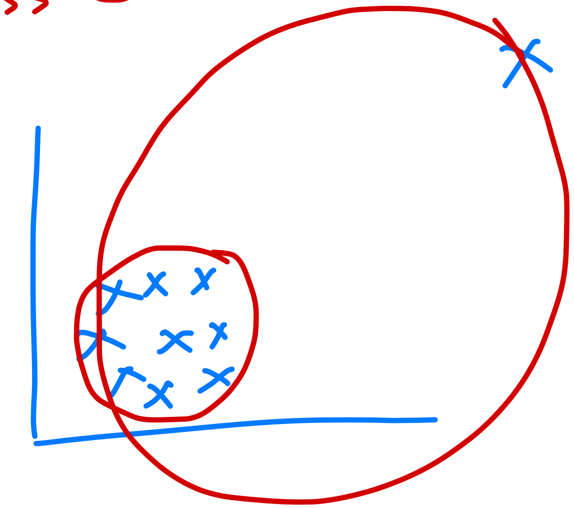


← describes the training set.

test data point x^*

$x^* \in \mathcal{X}$ or $x^* \notin \mathcal{X}$

{ outlier detection
{ anomaly detection
{ one-class classification



a = center of the circle

R = radius of the circle

$$\text{minimize} \quad R^2 + C \sum_{i=1}^{N} \varepsilon_i$$

$$\text{subject to:} \quad \|x_i - a\|_2^2 \leq R^2 + \varepsilon_i \quad \forall i$$

Exercise

$$\text{maximize} \quad \sum_{i=1}^{N} \alpha_i \underbrace{x_i^T x_i}_{k(x_i, x_i)} - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \underbrace{x_i^T x_j}_{k(x_i, x_j)}$$

$$\text{subject to:} \quad \sum_{i=1}^{N} \alpha_i = 1$$

$$C \geq \alpha_i \geq 0 \quad \forall i$$

x^*

$\|x^* - a\|_2^2 \leq R^2$

TRUE → Not outlier

False → Outlier