Multiclass Kernel Machines

$$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N} \qquad y_i \in \{1, 2, 3, \dots, K\}$$

one-versus-all $\Rightarrow$

1^+ vs $\{2, 3, 4, \dots, K\}^-$ $\Rightarrow$ SVM_1

2^+ vs $\{1, 3, 4, \dots, K\}^-$ $\Rightarrow$ SVM_2

3^+ vs $\{1, 2, 4, \dots, K\}^-$ $\Rightarrow$ SVM_3

$\vdots$

K^+ vs $\{1, 2, 3, \dots, K-1\}^-$ $\Rightarrow$ SVM_K

# of classifiers = K

training set size = N

test data point

$\downarrow$

x^*

$f_1(x^*)$ $f_2(x^*)$ $\underline{\underline{f_3}}(x^*)$ ---- $f_K(x^*)$ $\Rightarrow$ pick the maximum one

one-versus-other

1^+ vs 2^- $\Rightarrow$ SVM_{1vs2}

1^+ vs 3^- $\Rightarrow$ SVM_{1vs3}

$\vdots$

$(K-1)^+$ vs K^- $\Rightarrow$ $SVM_{(K-1)vsK}$

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

training set size $= N\frac{2}{K}$

$\Uparrow$

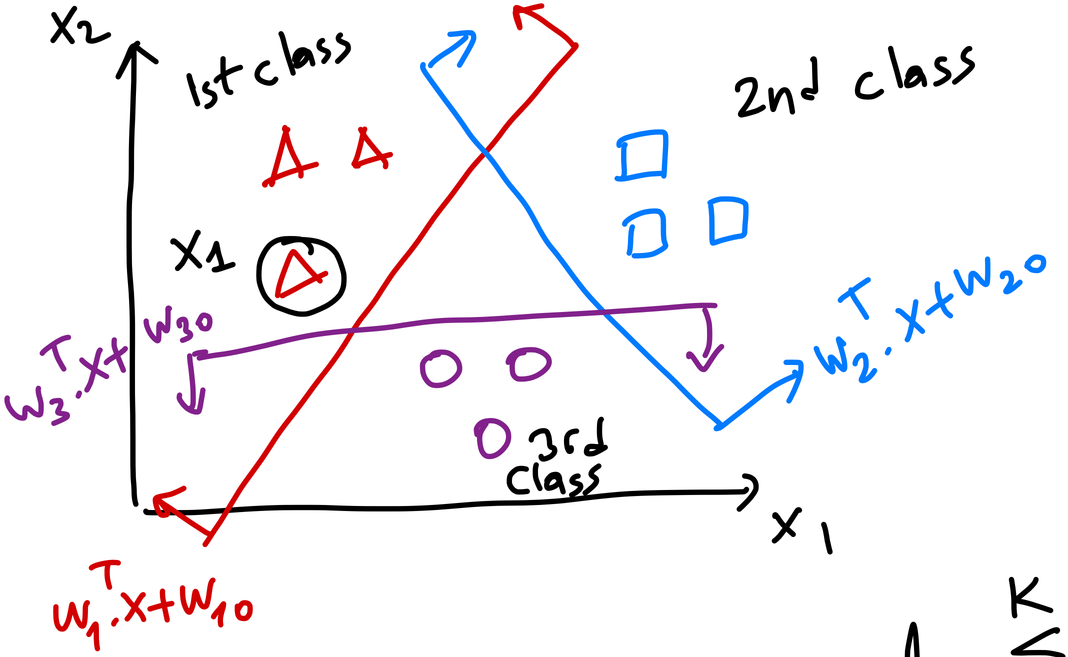
assuming classes are of the same size

x^*

$f_{1vs2}(x^*)$ $f_{1vs3}(x^*)$ .... $f_{(K-1)vsK}(x^*)$

$\Downarrow$

pick the one with maximum # of wins.



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

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

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

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

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

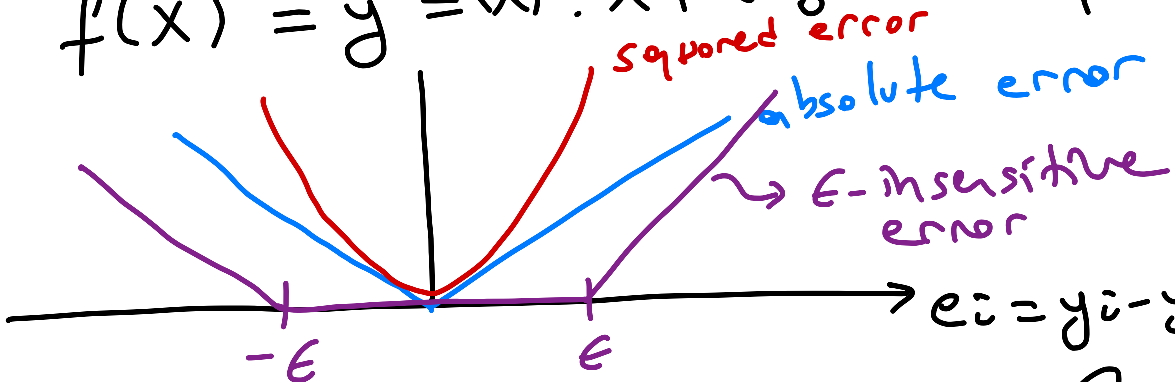
\# of decision variables $= (D+1).K + N(K-1)$

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

Kernel Machines for Regression

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

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



if $\hat{y}_i > y_i \Rightarrow$ overestimation

if $\hat{y}_i < y_i \Rightarrow$ underestimation

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

# of decision variables $= D + 1 + 2N$

# of constraints $= 2N$

PRIMAL:

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

$$\text{subject to: } \alpha_i^+ \; [\, y_i - \underbrace{[w^T . x_i + w_0]}_{\hat{y}_i} \leq \epsilon + \xi_i^+ \quad \forall i$$

$$\alpha_i^- \; [\, \underbrace{[w^T . x_i + w_0]}_{\hat{y}_i} - y_i \leq \epsilon + \xi_i^- \quad \forall i$$

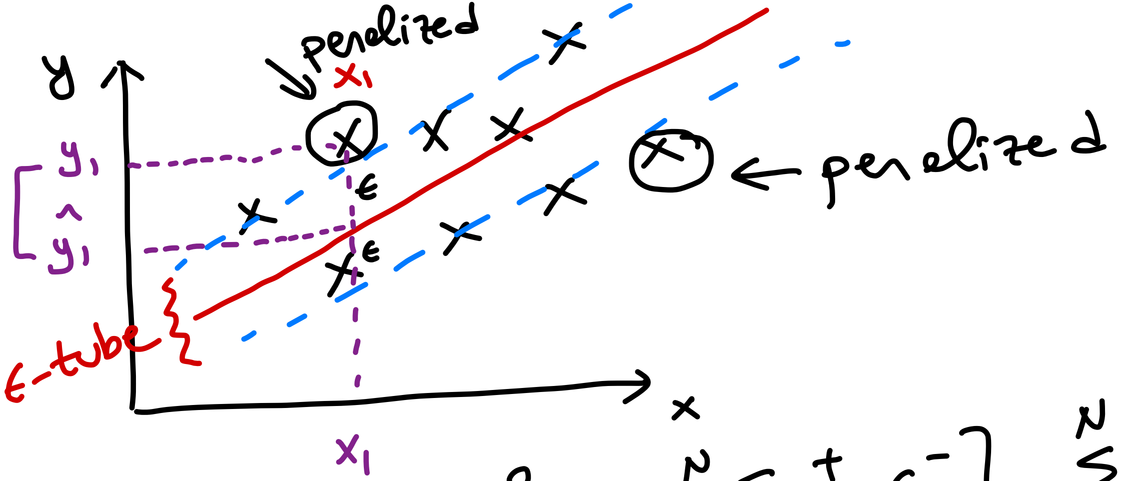
$$\beta_i^+ \; [\, \xi_i^+ \geq 0 \quad \forall i$$

$$\beta_i^- \; [\, \xi_i^- \geq 0 \quad \forall i$$

$y_i = 5 \quad \hat{y}_i = 8 \quad \epsilon = 2$

$y_i - \hat{y}_i = -3 \leq 2$

$\hat{y}_i - y_i = +3 \cancel{\leq} 2$



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

$$-\sum_{i=1}^{N}\alpha_i^-\left[-w^T x_i - w_0 + y_i + \epsilon + \varepsilon_i^-\right] - \sum_{i=1}^{N}\beta_i^+\varepsilon_i^+ - \sum_{i=1}^{N}\beta_i^-\varepsilon_i^-$$

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

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

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

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

DUAL:

Exercise: Find the dual formulation.

$$\text{maximize} \sum_{i=1}^{N} y_i[\alpha_i^+ - \alpha_i^-] - \epsilon \sum_{i=1}^{N} (\alpha_i^+ + \alpha_i^-) - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) \underline{\underline{x_i^T . x_j}}$$

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

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

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

We can replace $x_i^T . x_j$ with $k(x_i, x_j)$

\# of decision variables $= 2.N$

\# of constraints $= 1$

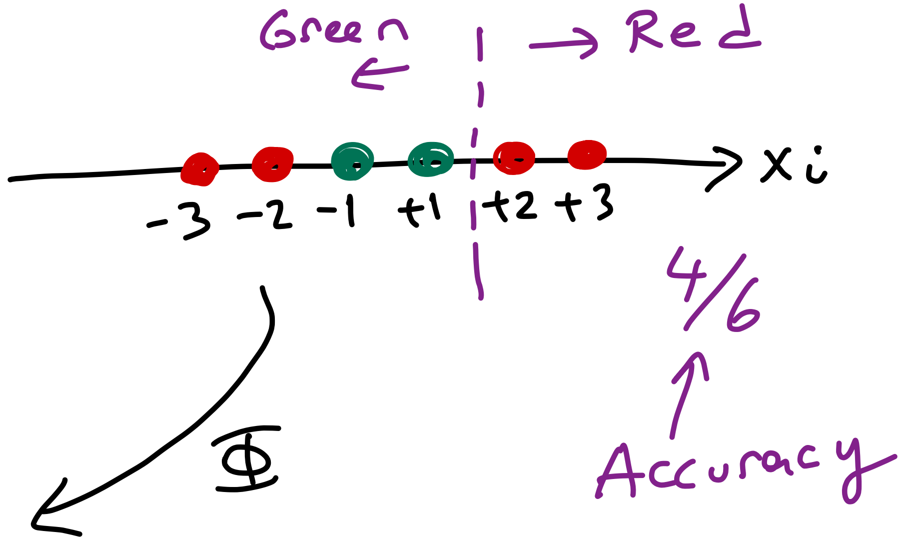
test data point $\rightarrow x^*$

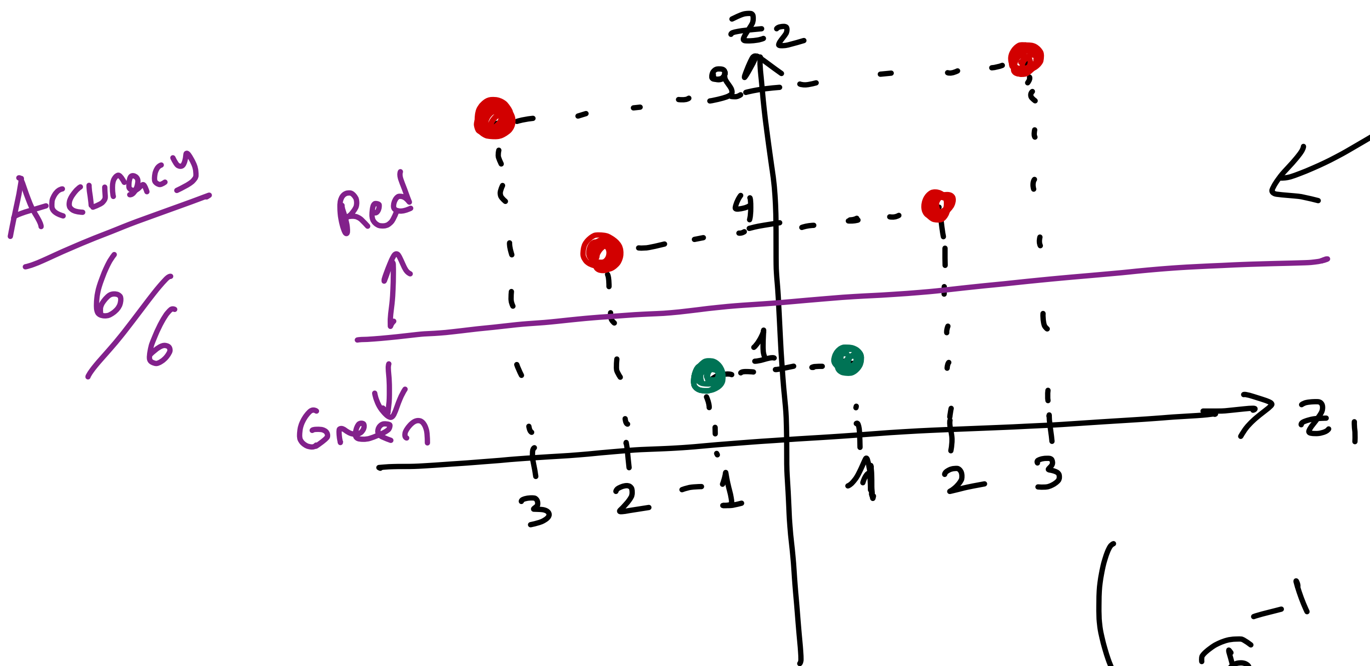
$$\Rightarrow f(x^*) = \underline{\underline{w}}^T . x^* + w_0 = \sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-) . \underbrace{x_i^T . x^*}_{k(x_i, x^*)} + w_0$$

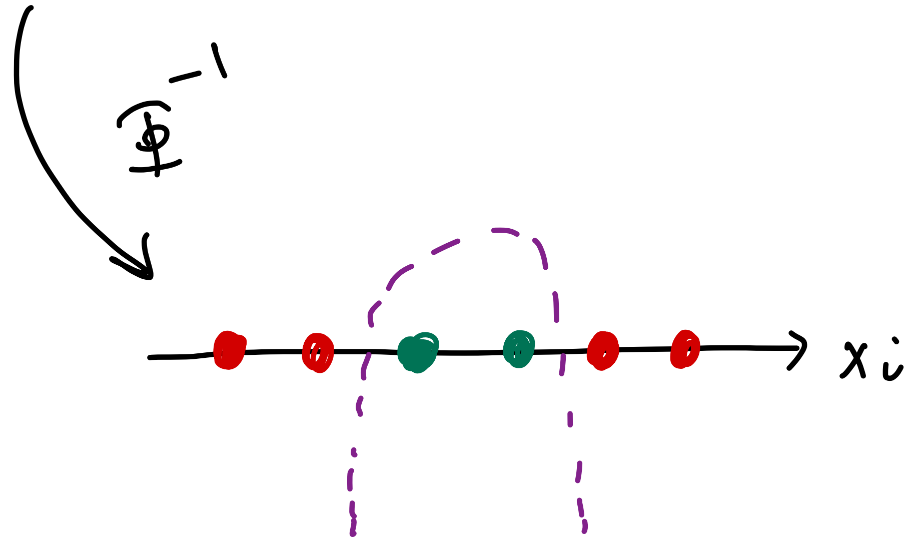
$$\text{extracting features} \leftarrow \Phi(x_i)^T . \Phi(x^*)$$

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

$D = 1$ $\qquad$ $\Theta = 2$





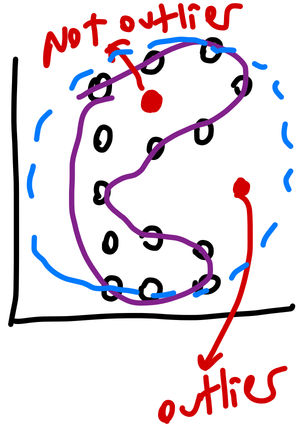


One-Class Kernel Machines

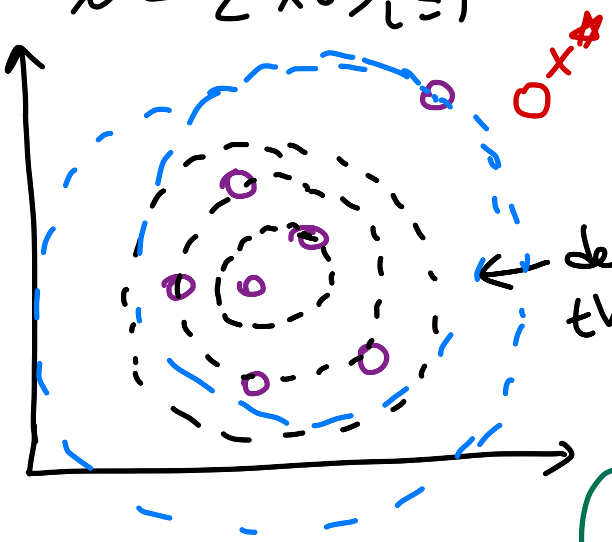
$\mathcal{X} = \{x_i\}_{i=1}^{N}$

test data point x^*

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



outlier detection
anomaly detection
one-class classification



describes the training set

a = center of the circle

R = radius of the circle

PRIMAL:

minimize $R^2 + C \sum_{i=1}^{N} \epsilon_i$

subject to: $\underbrace{||x_i - a||_2^2}_{\text{Euclidean distance between } x_i \text{ \& } a.} \leq R^2 + \epsilon_i \quad \forall i$

$\epsilon_i \geq 0 \quad \forall i$

Exercise

DUAL:

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

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

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

$\underline{x^*}$

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

True (Not outlier) False (outlier)