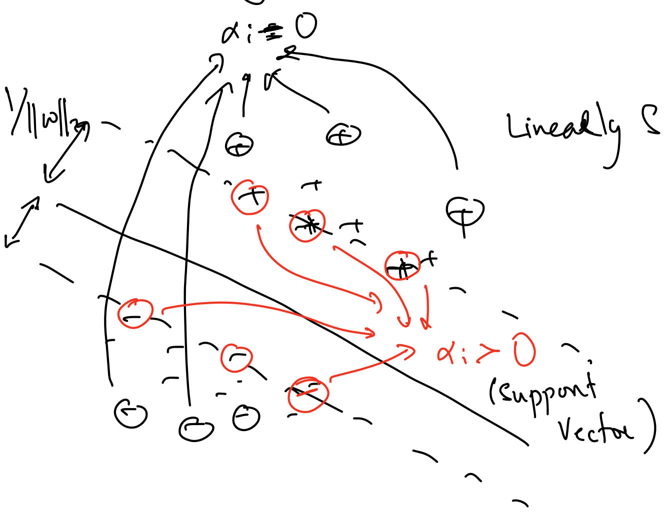
Non-Linearly Separable SVMs (Support Vector Machines)

& Kernel SVMs



Linearly Separable SVMs:

SVM Primal:

$$\min_{w, w_0} \quad \frac{1}{2}\|w\|_2^2$$

$$1 - y_i(w^T x_i + w_0) \le 0, \quad i = 1, 2, \dots N$$

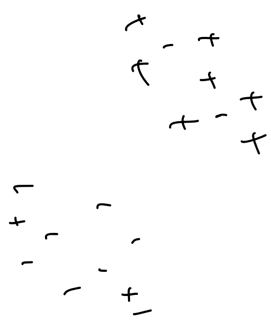
SVM Dual:

$$\max_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \right)$$

$$\text{st} \quad \alpha_i \ge 0, \quad i = 1, 2, \dots N$$

$$\sum \alpha_i y_i = 0$$

$$\Rightarrow \boxed{w^* = \sum_{i=1}^{N} \alpha_i y_i x_i}$$



Non-Linearly Separable SVM

Primal:

$$\min_{w, w_0, \xi} \quad \frac{1}{2}\|w\|^2$$

$$y_i(w^T x_i + w_0) \ge 1 - \xi_i, \quad i = 1, 2, \dots N$$

$$\xi_i \ge 0$$

$$\sum_{i=1}^{N} \xi_i \le \text{constant}$$

Primal:

$$\min_{w, w_0, \xi} \quad \frac{1}{2}\|w\|_2^2 + \gamma \left(\sum_{i=1}^{N} \xi_i \right)$$

$$\text{st} \quad 1 - \xi_i - y_i(w^T x_i + w_0) \le 0, \quad i = 1, 2, \dots N$$

$$-\xi_i \le 0, \quad i = 1, 2, \dots N$$

Lagrangian

$$L(w, w_0, \xi, \alpha, \mu) = \frac{1}{2}\|w\|_2^2 + \gamma \left(\sum_{i=1}^{N} \xi_i \right) + \sum_{i=1}^{N} \alpha_i \left(1 - \xi_i - y_i(w^T x_i + w_0)\right) - \sum \mu_i \xi_i$$

$$\nabla_w L = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\nabla_{w_0} L = 0 \Rightarrow \boxed{\sum_{i=1}^{N} \alpha_i y_i = 0}$$

$$\nabla_{\xi} L = 0 \Rightarrow \boxed{r - \alpha_i - \mu_i = 0}, \quad i = 1, 2, \ldots N$$

$$\alpha_i = r - \mu_i$$

Substituting back into the Lagrangian:

$$\frac{1}{2}\left\|\sum_{i=1}^{N}\alpha_i y_i x_i\right\|_2^2 + \underbrace{\sum_{i=1}^{N}\xi_i(r - \alpha_i - \mu_i)}_{=0} + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i y_i x_i^T\left(\sum_{i=1}^{N}\alpha_i y_i x_i\right) - \underbrace{\sum_{i=1}^{N}\alpha_i y_i w_0}_{=0}$$

$$-\frac{1}{2}\left\|\sum_{i=1}^{N}\alpha_i y_i x_i\right\|_2^2 + \sum_{i=1}^{N}\alpha_i$$

Dual Problem:

$$\max_{\alpha, \mu} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

st $\alpha_i \geq 0$, $\mu_i \geq 0$, $i = 1, 2, \ldots N$ $\quad \alpha_i = r - \mu_i \rightarrow \alpha_i \leq r$

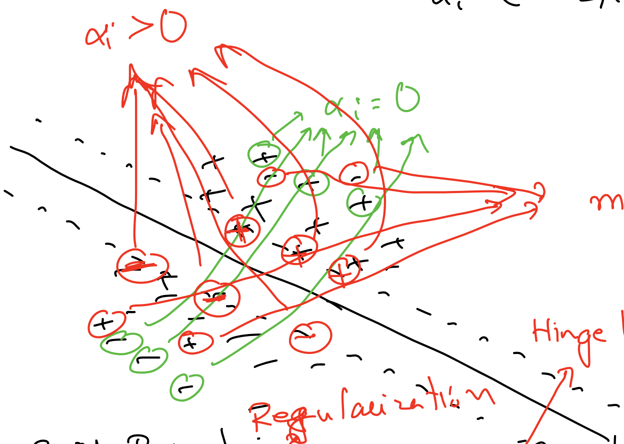
$$\sum \alpha_i y_i = 0$$

$$\boxed{\begin{array}{l}\text{SVM Dual Problem:} \\ \max\limits_{\alpha} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ \text{st } 0 \leq \alpha_i \leq r, \quad i = 1, 2, \ldots N \\ \sum_{i=1}^{N}\alpha_i y_i = 0\end{array}}$$

Complementary Slackness Conditions:

$$\alpha_i\left(1 - \xi_i - y_i(w^T x_i + w_0)\right) = 0, \quad i = 1, 2, \ldots N$$

$$\mu_i \xi_i = 0, \quad i = 1, 2, \ldots N.$$



misclassified points, $\xi_i > 0$, $\mu_i = 0$

$$\boxed{\alpha_i = r - \mu_i = r}$$

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i$$

Hinge Loss

Regularization

SVM Primal:

$$\boxed{\begin{array}{l}\min\limits_{w, w_0, \xi} \frac{1}{2}\|w\|^2 + r\sum_{i=1}^{N}\xi_i \\ \text{st } \xi_i \geq 0, \ i = 1, 2, \ldots N \\ 1 - \xi_i - y_i(w^T x_i + w_0) \leq 0, \ i = 1, 2, \ldots N\end{array}}$$

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

SVM Dual:

$$\boxed{\begin{array}{l}\max\limits_{\alpha} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ \text{st } 0 \leq \alpha_i \leq r, \ i = 1, 2, \ldots N \\ \sum_{i=1}^{N}\alpha_i y_i = 0\end{array}}$$

Define $X = [x_1\ x_2 \cdots x_N]$, $Y = \text{diag}([y_1\ y_2 \cdots y_N])$

$(X^TX)_{ij} = x_i^T x_j$, $e = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$, $e^T = [1\ 1 \cdots 1]$, $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}$

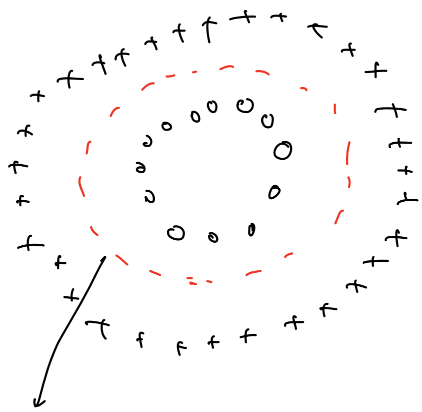
Gram matrix

Given the above notation, the dual SVM can be written as

$$\max_{\alpha} \; e^T\alpha - \frac{1}{2}\alpha^T \overbrace{Y^T X^T X Y}^{Q} \alpha = e^T\alpha - \frac{1}{2}\alpha^T Q \alpha$$

$$\underline{0} \le \underline{\alpha} \le \gamma \underline{e}$$

$$\alpha^T y = 0$$



$x_1^2 + x_2^2 = r$

not linear in x

But it is linear in $\phi(x)$

ϕ : inputspace $\rightarrow$ feature space

$\mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ $\quad d' > d$

feature space

$x \in \mathbb{R}^d$

$d = 2$, $\quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \phi(x) = \begin{bmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix}$, $d' = 6$

input space

$x' = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \rightarrow \phi(x') = \begin{bmatrix} 1 \\ \sqrt{2}x_1' \\ \sqrt{2}x_2' \\ (x_1')^2 \\ (x_2')^2 \\ \sqrt{2}x_1'x_2' \end{bmatrix}$

$$\phi(x)^T\phi(x') = [1\ \sqrt{2}x_1\ \sqrt{2}x_2\ x_1^2\ x_2^2\ \sqrt{2}x_1x_2] \begin{bmatrix} 1 \\ \sqrt{2}x_1' \\ \sqrt{2}x_2' \\ (x_1')^2 \\ (x_2')^2 \\ \sqrt{2}x_1'x_2' \end{bmatrix}$$

$$= 1 + 2x_1x_1' + 2x_2x_2' + x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2x_1x_2x_1'x_2'$$

$$= (1 + x_1x_1' + x_2x_2')^2$$

$$= (1 + x^T x')^2 \longrightarrow \text{Polynomial Kernel of degree 2}$$

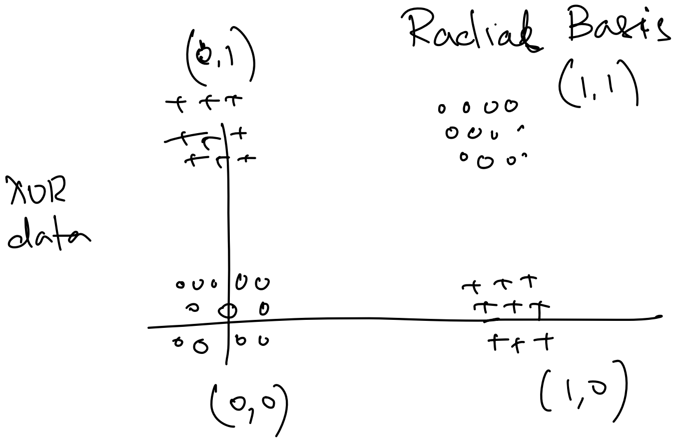
$$= K(x, x')$$

Kernel 'Trick' $\quad K(x, x') = \phi(x)^T \phi(x')$

Examples of kernel functions

m^{th} degree polynomial, $K(x, x') = (1 + x^T x')^m$

Radial Basis or Gaussian Kernel, $K(x, x') = e^{-\|x - x'\|^2 / c}$



LIBLINEAR

LIBSVM (Kernel functions)

Regression : Loss Function + Regularization

Ridge (Squared L_2)

Lasso (L_1)

Classification : Loss Function + Regularization (L_2 / L_1)

$$\ell_{logistic} = \log(1 + e^{-y\hat{y}})$$

SVMs : $\ell_{hinge} = \max(0, 1 - y\hat{y})$

If x_i is correctly classified $\xi_i = 0$

If not, ξ_i