Non-linearly Separable SVMs (Support Vector Machines) & Kernel SVMs

Recall: Linearly Separable SVM:

SVM Primal: $\min_{w} \frac{1}{2}\|w\|_2^2$

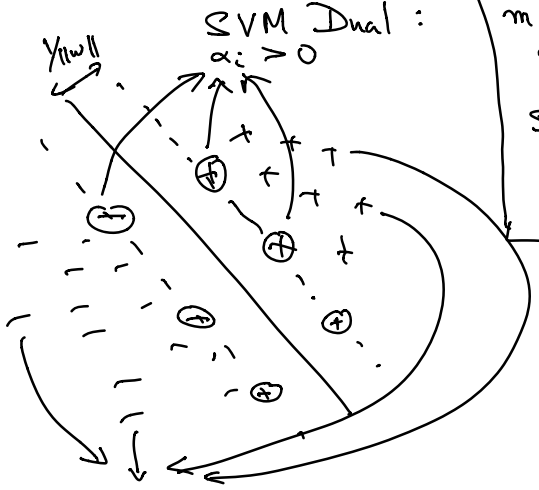
$1 - y_i(w^T x_i + w_0) \le 0$, for $i = 1, 2, \ldots N$

SVM Dual:

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

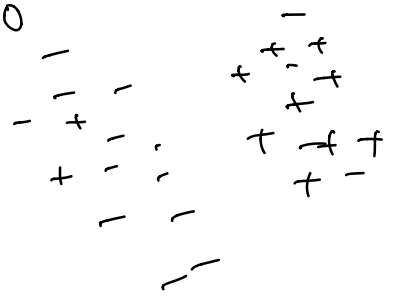
s.t. $\alpha_i \ge 0$, $i = 1, 2, \ldots N$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$



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

Non-linearly Separable SVM



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

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

$\xi_i \ge 0$

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

Primal:

$$\min_{w, w_0, \xi} \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{N} \xi_i$$

s.t. $1 - \xi_i - y_i(w^T x_i + w_0) \le 0$, $i = 1, 2, \ldots N$ — ①

$\xi_i \ge 0$, $i = 1, 2, \ldots N$ — ②

$-\xi_i \le 0$

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

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

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

$\nabla_{\xi_i} L = 0 \Rightarrow \gamma - \alpha_i - \mu_i = 0 \Rightarrow \alpha_i = \gamma - \mu_i$

Substituting back into Lagrangian:

$$\frac{1}{2}\left\|\sum_{i=1}^{N} \alpha_i y_i x_i\right\|^2 + \sum_{i=1}^{N} \xi_i(\gamma - \alpha_i - \mu_i) + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y_i x_i^T \left(\sum_{k=1}^{N} \alpha_k y_k x_k\right)$$

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

Dual Problem: $\max_{\alpha, \mu} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$

such that $0 \le \alpha_i, \mu_i \ge 0 \quad (\alpha_i = \gamma - \mu_i)$

$\sum_{i=1}^{N} \alpha_i y_i = 0$

$0 \le \alpha_i \le \gamma, \quad i = 1, 2, \ldots N$

SVM Dual:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

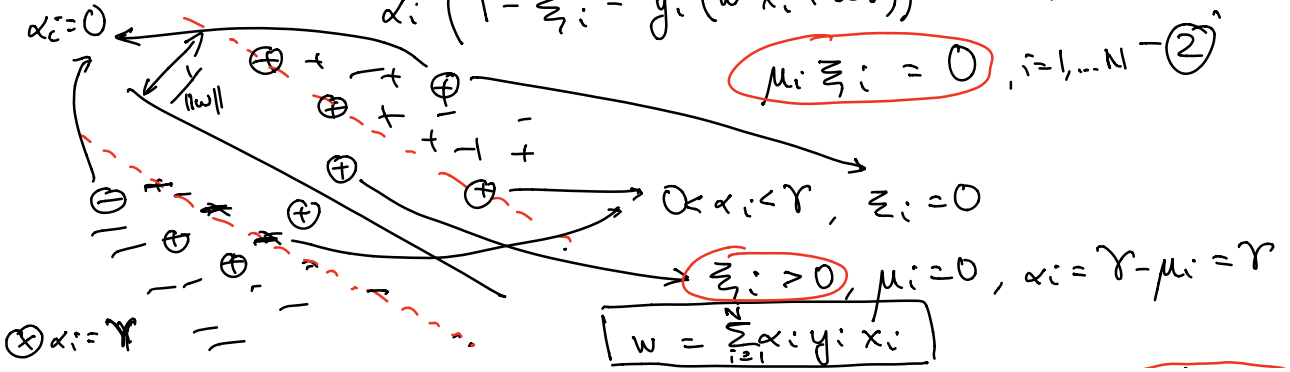
$$0 \le \alpha_i \le \gamma, \quad i = 1, 2, \ldots N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Complementary Slackness Conditions:

$$\alpha_i \left(1 - \xi_i - y_i(w^T x_i + w_0)\right) = 0, \quad i = 1, \ldots N \quad (1)'$$

$$\mu_i \xi_i = 0, \quad i = 1, \ldots N \quad (2)'$$



SVM Primal:

Regularization, hinge loss

$$\min_{w, w_0} \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{N} \xi_i$$

$$\text{s.t.} \quad \xi_i \geq 0, \quad i = 1, 2, \ldots N$$

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

SVM Dual:

$K(x_i, x_j)$

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

$$\text{s.t.} \quad 0 \leq \alpha_i \leq \gamma, \quad i = 1, 2, \ldots N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

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

$(X^T X)_{ij} = x_i^T x_j$

↓

Gram Matrix

$$e = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$e^T = [1 \; 1 \ldots 1]$$

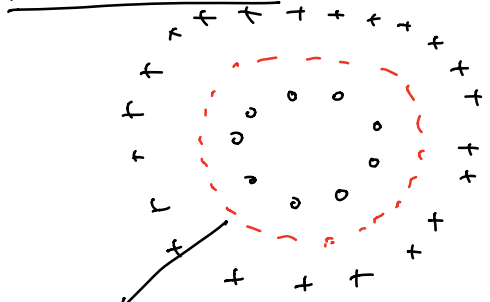
$$\max_{\alpha} \; e^T \alpha - \frac{1}{2} \alpha^T \underbrace{Y^T X^T X Y}_{Q} \alpha$$

$$0 \leq \alpha \leq \gamma e$$

$$\text{and } \alpha^T y = 0$$

LIBLINEAR

Kernel Methods



$x_1^2 + x_2^2 = r$

↓

not linear in x

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

$x \in \mathbb{R}^d \rightarrow$ quadratic feature map $\rightarrow$ feature space $\phi(x) \in \mathbb{R}^{d^2}$

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

input space

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

$$K(x,x') = (1 + x^T x')^2 \rightarrow \text{Polynomial kernel of degree 2}$$

$$= (1 + x_1 x_1' + x_2 x_2')^2$$

$$= 1 + 2x_1 x_1' + 2x_2 x_2' + 2x_1 x_2 x_1' x_2' + x_1^2 x_1'^2 + x_2^2 x_2'^2$$

$$= \begin{bmatrix} 1 & \sqrt{2}x_1 & \sqrt{2}x_2 & x_1^2 & x_2^2 & \sqrt{2}x_1 x_2 \end{bmatrix} \begin{pmatrix} 1 \\ \sqrt{2}x_1' \\ \sqrt{2}x_2' \\ x_1'^2 \\ x_2'^2 \\ \sqrt{2}x_1' x_2' \end{pmatrix}$$

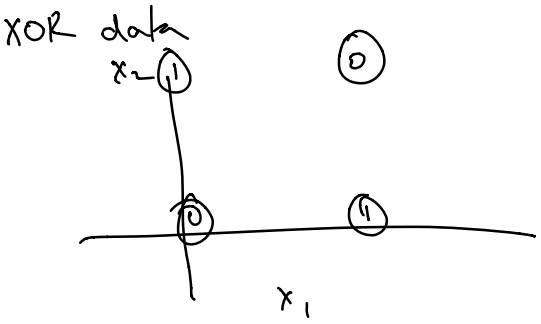
$$K(x,x') = \phi(x)^T \phi(x')$$

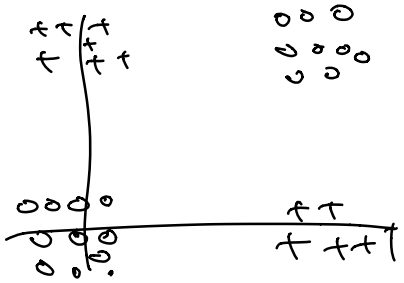
Kernel "trick"

Examples of kernel functions

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

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





LIBSVM

Difficulty in choosing appropriate kernel function

Regression : Loss function + Regularization

Ridge (squared L_2)

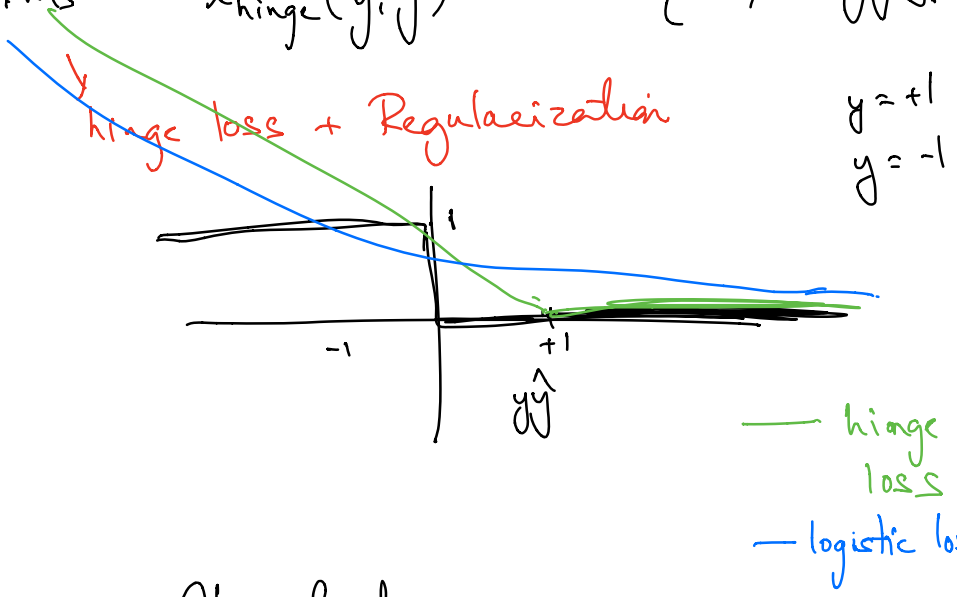
Lasso (L_1)

Classification : Loss function + Regularization

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

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

hinge loss + Regularization



Regression or Classification:

Loss + Regularization

Regression : L_2^2 loss or L_1 loss

Classification : Logistic loss or hinge loss, squared hinge loss