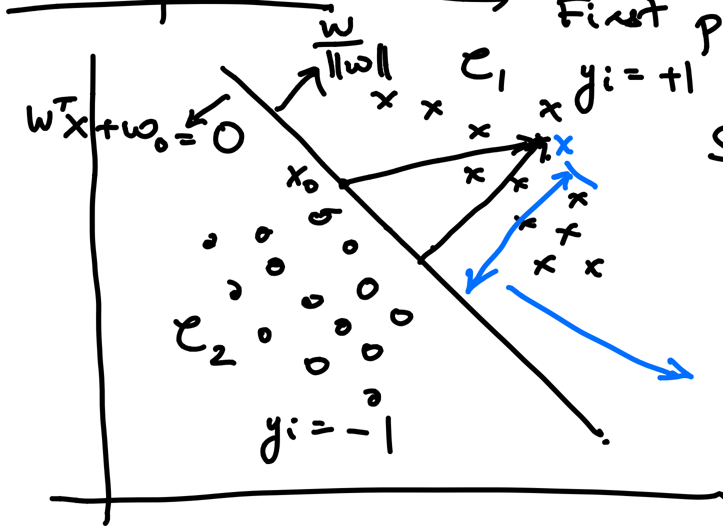
Mar 6, 2020

Perceptron

→ First proposed in 1962



Signed distance of x to hyperplane:

$$\frac{w^T}{\|w\|}(x - x_0)$$

$$= \frac{w^T x - w^T x_0}{\|w\|}$$

$$= \frac{w^T x + w_0}{\|w\|}$$

(because $w^T x_0 + w_0 = 0$)

For class $\mathcal{C}_1$, $w^T x_i + w_0 > 0$ for all points x_i in $\mathcal{C}_1$ that are correctly classified

For class $\mathcal{C}_2$, $w^T x_i + w_0 < 0$ for all points x_i in $\mathcal{C}_2$ that are correctly classified

For correctly classified points,

$$y_i (w^T x_i + w_0) > 0,$$

For misclassified points:

$$y_i (w^T x_i + w_0) < 0.$$

Perceptron Criterion

$$\min_{D(w,w_0)} D(w,w_0) = \left[-\sum_{i \in \mathcal{M}} y_i (w^T x_i + w_0) \right]$$

→ $\mathcal{M}$ indexes the misclassified points

$$\nabla_w D(w,w_0) = -\sum_{i \in \mathcal{M}} y_i x_i = -\left(\sum_{i \in C_1 \cap \mathcal{M}} x_i - \sum_{j \in C_2 \cap \mathcal{M}} x_j \right)$$

$$\nabla_{w_0} D(w,w_0) = -\sum_{i \in \mathcal{M}} y_i = -(N_1 - N_2)$$

where N_i is the number of misclassified points in C_i

Perceptron Update Rule

$$\begin{bmatrix} w \\ w_0 \end{bmatrix} \leftarrow \begin{bmatrix} w \\ w_0 \end{bmatrix} + \eta \begin{bmatrix} y_i x_i \\ y_i \end{bmatrix}$$

Stochastic Gradient Descent: look at x_i, and if x_i is misclassified go in direction of negative gradient (but only contribution to gradient by x_i).

Perceptron pseudocode

Start with some w, w_0

Repeat

for $i = 1$ to n do

if $y_i(w^T x_i + w_0) < 0$ then

$w \leftarrow w + \eta y_i x_i$

$w_0 \leftarrow w_0 + \eta y_i$

end if

end for

until there are no mistakes (misclassifications) within the for loop

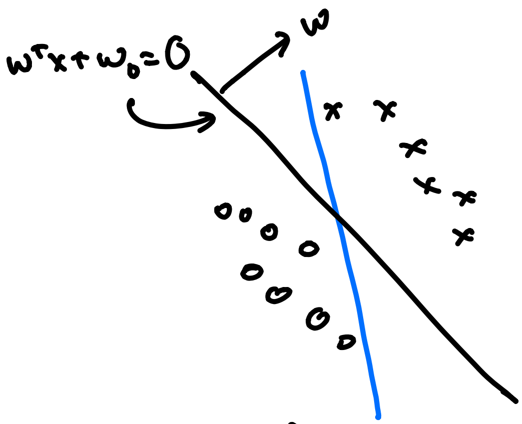
Perceptron algorithm is guaranteed to find a separating hyperplane <u>if</u> the data is linearly separable.

Drawbacks of Perceptron:

1. If the data is linearly separable, the hyperplane output depends on the order in which points (data) is presented to the algorithm
2. Number of iterations might be large
3. If classes are not linearly separable, then the algorithm will not converge - cycles can

develop that are not easy to detect.

Support Vector Machines



Signed distance of x to hyperplane:

$$\frac{w^T x - w^T x_0}{\|w\|} = \frac{w^T x + w_0}{\|w\|}$$

$$\frac{y_i (w^T x_i + w_0)}{\|w\|}$$ – Distance of each training point x_i to the hyperplane

Suppose we put requirement that each of these distances is greater than C

$$\frac{y_i (w^T x_i + w_0)}{\|w\|} \geq C$$

$$\Rightarrow \quad y_i (w^T x_i + w_0) \geq C \cdot \|w\|$$

In linear support vector machines, goal is to maximize C.

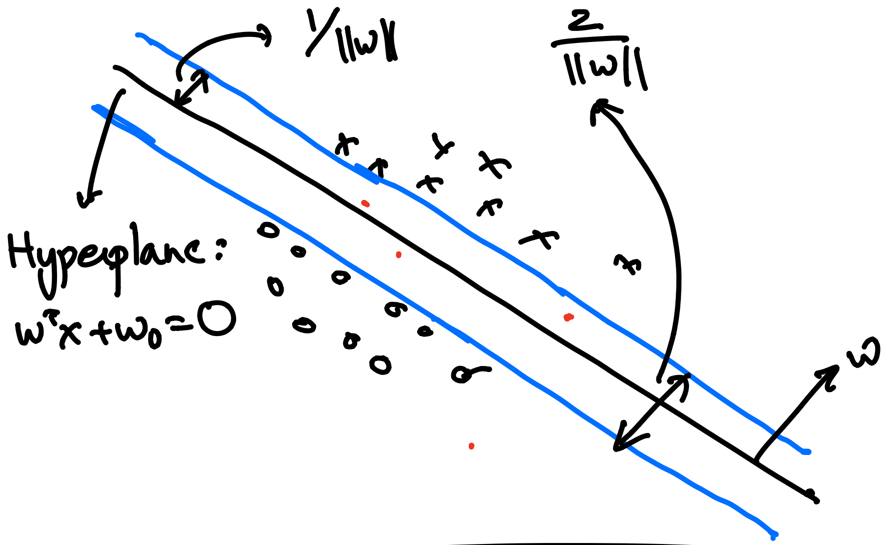
$$\boxed{\begin{array}{l} \underset{w, w_0}{\text{Maximize}} \quad C \\ \text{such that} \quad y_i (w^T x_i + w_0) \geq C \|w\| \end{array}}$$

Fix $C\|w\| = 1$ (since I can arbitrarily scale w & w_0)

$\Rightarrow C = \frac{1}{\|w\|}$

Primal SVM Problem:

$$\boxed{\begin{aligned} &\underset{w,w_0}{\text{Minimize}} \ \|w\|^2 \\ &\text{such that} \ y_i(w^T x_i + w_0) \geq 1, \ i=1,2\ldots N \end{aligned}}$$



General constrained optimization problem

$$\left.\begin{aligned} &\min_x \ f_0(x) \\ &\text{st} \ \ f_i(x) \leq 0, \ i=1,2,\ldots m \\ &\qquad h_i(x) = 0, \ i=1,2,\ldots p \end{aligned}\right\} \text{Primal} \quad -(1)$$

Here $x \in \mathbb{R}^n$ (i.e., $f_i : \mathbb{R}^n \to \mathbb{R}$, $h_i : \mathbb{R}^n \to \mathbb{R}$)

Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

where λ_i is Lagrange multiplier for i-th inequality constraint

& ν_i " " " " " equality "

λ & ν are also dual variables

Lagrange dual function:

$$g(\lambda,\nu) = \inf_x L(x,\lambda,\nu) = \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

λ & ν are dual feasible if $\lambda \geq 0$ & $g(\lambda,\nu) > -\infty$

Fact 1. $g(\lambda,\nu) \leq p^*$ for any dual feasible λ, ν (where p^* is optimal value of (1))

$p^* = f_0(x^*)$

Fact 2. If $\exists$ dual feasible λ^*, ν^* $(\lambda^* \geq 0)$ & primal feasible x^* st $g(\lambda^*,\nu^*) = p^* = f_0(x^*)$

then strong duality is said to hold

Dual Problem : $\boxed{\begin{array}{ll} \max\limits_{\lambda,\nu} & g(\lambda,\nu) \\ \text{st} & \lambda \geq 0 \end{array}}$

Suppose strong duality holds for λ^*, ν^*, x^*

Then $f_0(x^*) = g(\lambda^*, \nu^*) = \inf_x L(x, \lambda^*, \nu^*)$

$$= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right)$$

$$\leq f_0(x^*) + \underbrace{\sum_{i=1}^{m} \underset{\geq 0}{\lambda_i^*} \underset{\leq 0}{f_i(x^*)}}_{\leq 0} + \sum_{i=1}^{p} \underbrace{\nu_i^* h_i(x^*)}_{=0}$$

$$\leq f_0(x^*)$$

So, the above two inequalities must hold with equality

① x^* must be minimizer of $L(x, \lambda^*, \nu^*)$

② $\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0 \Rightarrow \lambda_i^* f_i(x^*) = 0, \; i = 1, 2, \ldots m$

if $\lambda_i^* > 0$, then $f_i(x^*) = 0$

if $f_i(x^*) < 0$, then $\lambda_i^* = 0$

Karush Kuhn Tucker

KKT Conditions : Provide certificate of optimality when problem ① is convex

KKT conditions for x^* & (λ^*, ν^*) :

Primal Feasible
$$\begin{cases} f_i(x^*) \leq 0 & , \; i = 1,2,\ldots m \\ h_i(x^*) = 0 & , \; i = 1,2,\ldots p \end{cases}$$

Dual Feasible : $\lambda_i^* \geq 0 \;,\; i = 1,2,\ldots m$

Complementary Slackness $\rightarrow \lambda_i^* f_i(x^*) = 0 \;,\; i = 1,2,\ldots m$

$$x^* = \arg\min_x L(x, \lambda^*, \nu^*)$$

$$\rightarrow \nabla f_0(x^*) + \sum \lambda_i \nabla f_i(x^*) + \sum \nu_i^* \nabla h_i(x^*) = 0$$

SVM Problem :

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

st $y_i(w^T x_i + w_0) \geq 1$, for $i = 1,2,\ldots N$

$\downarrow$

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

Lagrangian

$\downarrow$

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

$$\nabla_w L(w, w_0, \alpha) = 0 \Rightarrow w + \sum_{i=1}^{N} -\alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\nabla_{w_0} L(w, w_0, \alpha) = 0 \Rightarrow -\sum_{i=1}^{N} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$$

Dual function: $g(\alpha) = \inf_{w, w_0} L(w, w_0, \alpha)$

$$g(\alpha) = \frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i y_i x_i\right\|^2 + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i y_i x_i^T\left(\sum_{j=1}^{N}\alpha_j y_j x_j\right) - \cancel{\sum \alpha_i y_i w_0}$$

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

<u>SVM Dual</u>: $\max_{\alpha} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j$

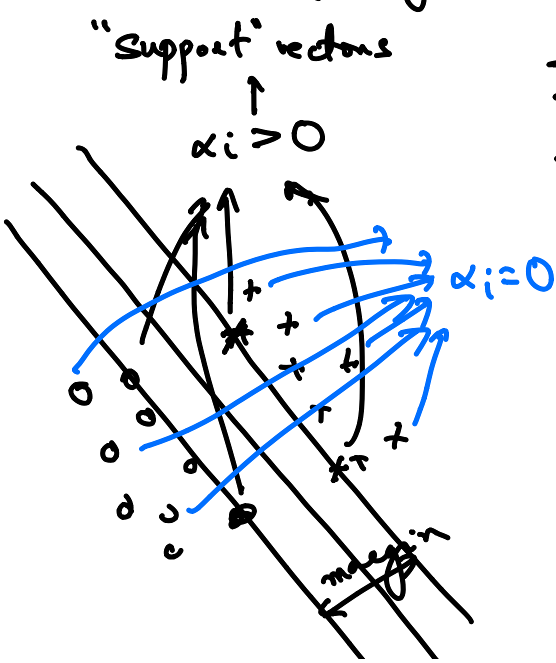
such that $\alpha_i \geq 0$, $i = 1, 2, \ldots N$

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

<u>Complementary Slackness</u>

At optimality, α_i, w, w_0 must satisfy:

$$\alpha_i\left(1 - y_i(w^T x_i + w_0)\right) = 0 \quad \text{for all } i$$



If $\alpha_i > 0$, then $y_i(w^T x_i + w_0) = 1$

If $y_i(w^T x_i + w_0) > 1$, then $\alpha_i = 0$

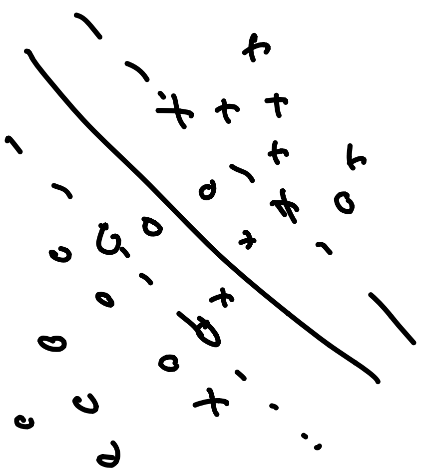
$$w = \sum_{i=1}^{N}\alpha_i y_i x_i$$

↓

linear combination of support vectors

Recall original separable SVM formulation:

$$\max_{w,w_0} C$$
$$\text{st} \quad y_i(w^T x_i + w_0) \geq C\|w\|$$



Relax above inequality to have a slack, i.e.,

$$y_i(x_i^T w + w_0) \geq C(1-\xi_i)\|w\|,$$
$$\xi_i \geq 0$$

$$C\|w\| = 1$$

Non-linearly separable SVM:

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

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

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

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

SVM Primal

$$\boxed{\begin{aligned} \min_{w,w_0,\xi} \quad & \frac{1}{2}\|w\|^2 + \gamma\left(\sum_{i=1}^{N} \xi_i\right) \\ & 1 - \xi_i - y_i(w^T x_i + w_0) \leq 0, \quad i = 1,2..N \\ & \xi_i \geq 0 \quad , i = 1,2,..N. \end{aligned}}$$