Support vector machines.

We have seen that mapping features into another space facilitates the learning of a linear classifier.

Eg: Let $X = [x_1 \ x_2]$

Define a $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^5$,

$$z = \phi(X) = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_2^2 \quad x_1 x_2]$$

if $g(x) = a_0 + a_1 x + a_2 x_2 + a_3 x_1^2 + a_4 x_2^2$
$+ a_5 x_1 x_2$

is a quadratic disc fn in $\mathbb{R}^2$.

$\phi(g(x)]$ is a linear disc in $\phi(x)$ space.

Two major issues with this idea.

1) If we want, p^{th} degree poly disc fn in the original feature space (R^m), then the transformed vector, z, has dim $O(m^p)$.

This results in a huge computational cost in both learning & inference.

2) Since we need $O(m^p)$ parameters than $O(m)$ paramters, we need large number of examples to achieve generalization.

SVM is designed to offer solution to both of these problems.

Let the training set be

$\{(x_i, y_i), i = 1, \ldots, n\}$, $x_i \in \mathbb{R}^m$, $y_i \in \{+1, -1\}$

To begin with assume training set is linearly separable.

$\exists\, W \in \mathbb{R}^m$ & $b \in R$ s.t

$W^T x_i + b > 0$ $\forall i$ s.t $y_i = +1$

$W^T x_i + b < 0$, $\forall i$ s.t $y_i = -1$

$W^T x + b = 0$ is a separating hyperplane

we know that there can be infinitely many separating hyperplanes

$\Rightarrow \quad \exists \; \epsilon > 0, \quad s.t$

$$w^T x_i + b \geq \epsilon, \quad \forall i \; s.t \; y_i = +1$$

$$w^T x_i + b \leq -\epsilon \quad \forall i, \quad s.t \; y_i = -1$$

Now, w can be scaled s.t

$$w^T x_i + b \geq +1 \quad \text{if } y_i = +1$$

$$w^T x_i + b \leq -1 \quad \text{if } y_i = -1$$

equivalently, $y_i (w^T x_i + b) \geq 1, \forall i.$

$\Rightarrow$ when training set is separable, any separating hyperplane can be scaled s.t

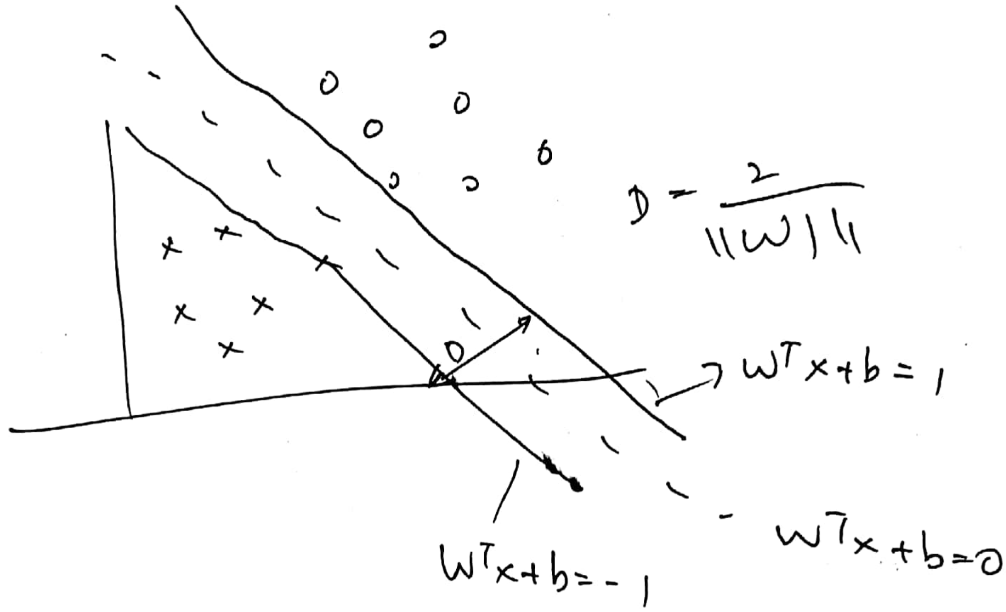
$$y_i (w^T x_i + b) \geq 1, \quad \forall i$$

$\Rightarrow \exists$ no training patterns b/w

$$w^T x + b = +1, \quad \& \quad w^T x + b = -1.$$

Distance b/w these two lines is : $\frac{2}{\|W\|}$.

This is called the margin of the separating hyper plane.



So a good hyperplane is one that maximize D.

$\therefore$ Find $W \in \mathbb{R}^m$, $b \in \mathbb{R}$ to

minimize $\frac{1}{2}W^TW$

s.t $y_i(W^Tx_i+b) \geq 1$, $i=1,\dots,n$.

The SVM optimization problem.

Find $w \in R^m$, $b \in R$ s.t

$$\text{minimize } \frac{w^T w}{2}$$

$$\text{subject to } y_i(w^T x_i + b) \geq 1, \; i = 1, \dots, n.$$

Quadratic cost with linear inequality constraints.

An overview of constrained optimization.

$$\text{minimize } f(x)$$
$$\text{subject to } a_j^T x + b_j \leq 0, \; j = 1, \dots, r.$$

$f : R^m \to R$, $a_j \in R^m$, $b_j \in R$, $j = 1, \dots, r$.

Any point $x \in \mathbb{R}^m$ is called feasible if

$a_j^T x + b_j \leq 0$, $j = 1, \dots, r$.

A constrained optimization problem is solved as follows: Define Lagrangian as follows:

$$L(x, y) = f(x) + \sum_{j=1}^{r} y_j \left(a_j^T x + b_j\right) \quad \text{— (1)}$$

If $f(x)$ is convex,

if $x_1, x_2 \in R^m \quad \forall \alpha \in (0,1)$

$$f\left(\alpha x_1 + (1-\alpha) x_2\right) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$$

$$f(x) = x^T x$$

Kuhn-Tucker conditions:

An optimization problem with f convex:

- Any x^* is a global minimum if & only if x^* is feasible & $\exists\, y_j^*,\ j=1,\dots,r$, s.t.

1. $\nabla_x L(x^*, y^*) = 0$

2. $y_j^* \geq 0,\ \forall j$

3. $y_j^*(a_j^T x^* + b_j) = 0,\ \forall j$ → Complementary slackness

Duality: For a given optimization,

Lagrangian is

$$L(x, y) = f(x) + \sum_{j=1}^{r} y_j (a_j^T x + b_j)$$

Here, $x \in \mathbb{R}^n$ & $y \in \mathbb{R}^r$.

(171).

Define Dual function: $q : R^r \to [-\infty, \infty]$ by

$$q(\mu) = \inf_x L(x, \mu)$$

The dual problem is

maximize $q(\mu)$

subject to $\mu_j \geq 0, \; j = 1, \ldots, r$

again a constrained opti problem.

over R^r & $\mu \in R^r$ are the varibles.