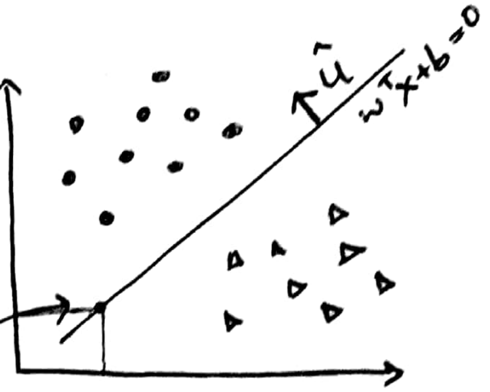
Perceptron and Logistic Regression can find a hyperplane that separates the data if the data is linearly separable. Then why SVM?

→ SVM finds the hyperplane with best separability (largest margin) and thus gives better GENERALIZATION performance.

$\boxed{\text{Goal}}$:

we want to find a line that not only separates 2 kinds of data but also maximizes the margin.

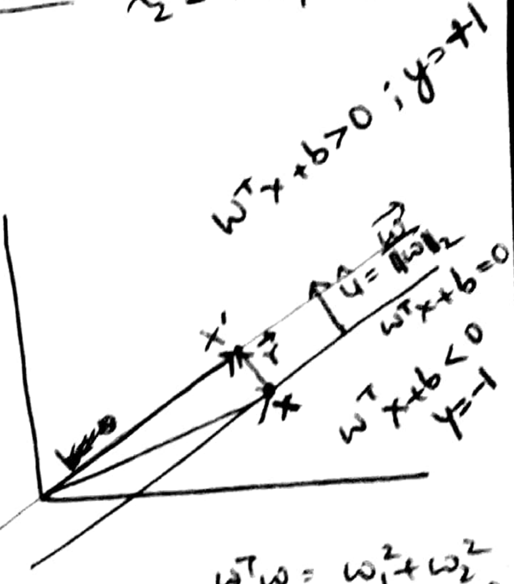


$$w^T x + b = 0$$

for $x \in R^2$ $\quad w_1 x_1 + w_2 x_2 + b = 0$

$$x_2 = \underbrace{\left(\frac{-w_1}{w_2}\right)}_{m} x_1 - \underbrace{\left(\frac{b}{w_2}\right)}_{\text{Intercept}}$$

thus form $\quad x_2 = m x_1 + c$



$$\hat{u} = \frac{\vec{w}}{\|\vec{w}\|_2}$$

$\vec{r} = \vec{x}' - \vec{x}$ — ①

let $r = r\hat{u}$

$= r \frac{\vec{w}}{\|w\|_2}$ — (*)

$w^T \vec{r} = w\vec{x}' - w\vec{x}$

$w^T w = w_1^2 + w_2^2$

$= \left(\sqrt{w_1^2 + w_2^2}\right)^2$

$= \|w\|^2$

$\cancel{w^T x} = -b \qquad \cancel{w^T x'} = 0$.

$w^T(\vec{x}' - \vec{x}) = b$ — ②

using ①

$w^T r = b \;\Rightarrow\; w^T \cdot r \frac{\vec{w}}{\|w\|} = b \;\Rightarrow\; r \frac{\|w\|^2}{\|w\|} = b$

$\vec{\gamma} = \vec{x}' - x$ — (a) $\quad \vec{\gamma} = \dfrac{\gamma \vec{w}}{\|w\|}$ — (b)

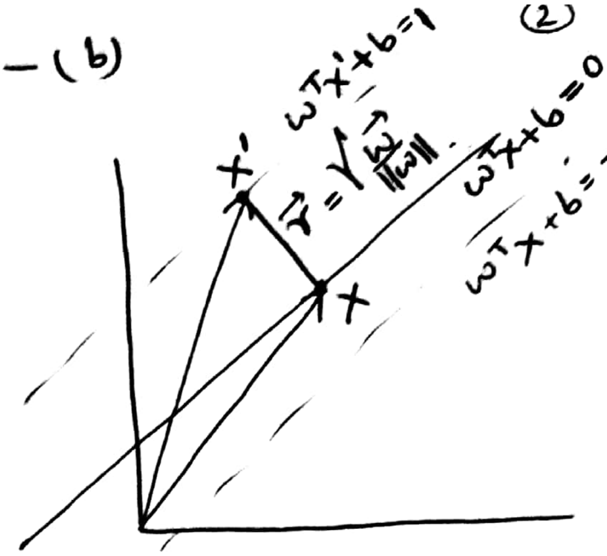
$w^T \vec{\gamma} = w^T \vec{x}' - w^T x$

$w^T \left(\dfrac{\gamma \vec{w}}{\|w\|} \right) = (1-b) - (b)$

$\gamma \dfrac{w^T \vec{w}}{\|w\|} = 1$

$\gamma \dfrac{\|w\|^2}{\|w\|} = 1$

$\therefore \gamma = \dfrac{1}{\|w\|}$



$w^T w = w_1^2 + w_2^2$
$= \left(\sqrt{w_1^2 + w_2^2}\right)^2$
$= \|w\|^2$

$$\boxed{\text{since margin} = 2\gamma = \frac{2}{\|w\|}}$$

We want to learn w, b such that for all points x_i for which $y_i = 1$ $\quad w^T x_i + b > 0$.
for all points x_i for which $y_i = -1$, $w^T x_i + b < 0$
and $\frac{2}{\|w\|}$ is as large as possible

$y_i(w^T x_i + b) > 0$

How to find w and b for the boundary?
Goal: Maximize margin while incurring 0 training error.

Maximize $\dfrac{2}{\|w\|}$ with 0 loss.

Minimize $\dfrac{\|w\|}{2}$ with 0 loss

or

Min $\dfrac{w^T w}{2}$ with 0 loss

Optimization Formulation

$$\underset{w,b}{\text{minimize}} \quad \frac{\|w\|^2}{2}$$

subject to $y_n(w^T x_n + b) \geq 1 \quad n = 1, 2, \ldots N$

or $1 - y_n(w^T x_n + b) \leq 0$

This is an optimization problem with N linear inequality constraint.

Let's turn the constrained optimization problem to unconstrained.

Approach

1. Add Lagrangian constant
2. solve Dual Optimization problem as in this case result will be equal to primal optimization problem as (a) function is convex
   (b) constraints are affine (linear)

$$d^* = \max_{\alpha, \beta : \alpha \geq 0} \min_{w} L(w, \alpha, \beta)$$

lagrangian α = inequality constant
β = equality constant

KKT conditions should be true

$$\underset{w,b,\alpha}{\text{minimize}} \; L(w, b, \alpha) = \frac{\|w\|^2}{2} + \sum_{x=1}^{N} \alpha_n \{1 - y_n(w^T x_n + b)\}$$

subject to $\alpha_n \geq 0 \quad \forall n$.

$$\frac{\partial}{\partial w} L(w, b, \alpha) = w - \sum_{x=1}^{N} \alpha_n y_n x_n$$

setting $\frac{\partial L(w, b, \alpha)}{\partial w} = 0$

gives $\boxed{w = \sum_{n=1}^{N} \alpha_n y_n x_n}$

solving Dual Optimization Problem!!

$$\frac{\partial}{\partial b} L(w,b,\alpha) = -\sum_{n=1}^{N} \alpha_n y_n$$

setting $\frac{\partial L}{\partial b}(w,b,\alpha) = 0$.

gives $\boxed{\sum_{n=1}^{N} \alpha_n y_n = 0}$

Substituting w in L

$$\frac{\|w\|^2}{2} + \sum_{n=1}^{N} \alpha_n \{1 - y_n(w^T x_n + b)\}$$

$$L(\alpha) = \frac{w^T w}{2} + \sum_{n=1}^{N} \alpha_n - \sum_{n=1}^{N} \alpha_n y_n w^T x_n - \underbrace{\left(\sum_{n=1}^{N} \alpha_n y_n\right) b}_{\text{poof!}}$$

$$= \frac{1}{2}\left(\sum_{n=1}^{N} \alpha_n y_n x_n\right)^T \left(\sum_{n=1}^{N} \alpha_n y_n x_n\right) + \sum_{n=1}^{N} \alpha_n - \sum \alpha_n y_n \left(\sum \alpha_n y_m x_m\right)^T x_n$$

Now we have a function of only α

Now we will $\max_{\alpha} L(\alpha)$.

$$L(\alpha) = \frac{1}{2} \sum_{n=1}^{N} \sum_{n=1}^{N} (\alpha_n \alpha_m y_n y_m x_m^T x_n) + \sum_{n=1}^{N} \alpha_n$$

$$- \sum_{n=1}^{N} \sum_{M=1}^{N} (\alpha_n \alpha_m y_n y_m x_m^T x_n).$$

$$\boxed{\max_{\alpha} L(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n} \sum_{m} \alpha_n \alpha_m y_n y_m x_m^T x_n}$$

subject to $\alpha_n > 0$ and $\sum \alpha_n y_n = 0$

This is a quadratic optimization problem

Offload to a Quadratic Programming solver.

Let $\alpha_1^*, \alpha_2^* \ldots \alpha_n^*$ be the solution of QP problem

$$\boxed{w = \sum_{n=1}^{N} \alpha_n y_n x_n}$$

Get w from here

$y_n\{w^T x_n + b\} -$
$b = \frac{1}{y_n} - w^T x$

We can get B using KKT condition #5

$$\sum_{n=1}^{N} \alpha_n \left(y_n \{ w^T x_n + b \} - 1 \right) = 0 \quad \text{and } \alpha_n \geq 0$$

Either α_n or $(y_n\{w^T x_n + b - 1)$ should be 0

If $\alpha_n > 0$ then $y_n\{w^T x_n + b\} - 1$ will be 0.
i.e. x_n is on the margin

If $\alpha = 0$. then $y_n\{w^T x_n + b\} - 1 \geq 0$
i.e. points are away from margin
and thus these points will not participate
as when we compute "w" only training
points which are on the margin participate

$$\boxed{w = \sum_{n=1}^{N} \alpha_n y_n x_n}$$

Only training examples that lie on the margin are relevant. These are called
Support Vectors

We only save value of non-zero α and b

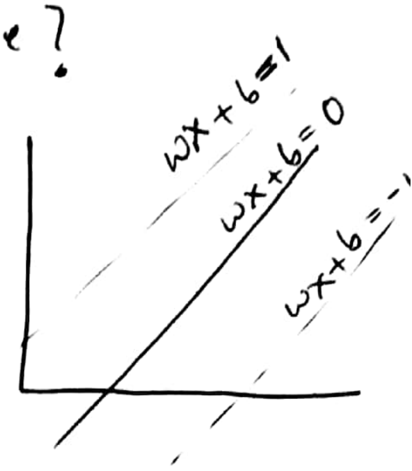
For a test point x^*.

$v = w^T x^* + b$.

$v = (\sum \alpha_n^* y_n x_n)^T x^* + b$.

No need to compute w. We directly use values of α & b.

What if data is not Linearly separable?



we relax the constraint

$$y_n(w^T x_n + b) \geq 1 - \xi_n \quad \forall n = 1 \dots N$$

ξ_n is called slack variable

Optimization Problem for Non-separable case.

$$\underset{w,b}{\text{minimize}} \; \cancel{\text{Maximize}} \quad f(w,b) = \|w\|^2 + C\sum_{n=1}^{1} \xi_n$$

subject to $y_n(w^T x_n + b) \geq 1 - \xi_n$ $\quad \xi_n \geq 0 \quad n = 1 \dots N$

C controls the impact of margin & margin error

Now support vectors are not just points which are on the margin but also those which are on the wrong side of the margin.

1. Points on the margin $(\xi_n = 0)$
2. Inside the margin but on correct side $(0 < \xi_n < 1)$
3. On the wrong side of the hyperplane $(\xi_n \geq 1)$

Support vectors will be more in number.