s.t $Ax \leq c$ & $x \geq 0$

For solving NLP there is no unique method to get the solution — Lagrangian multiplier approach is one approach.

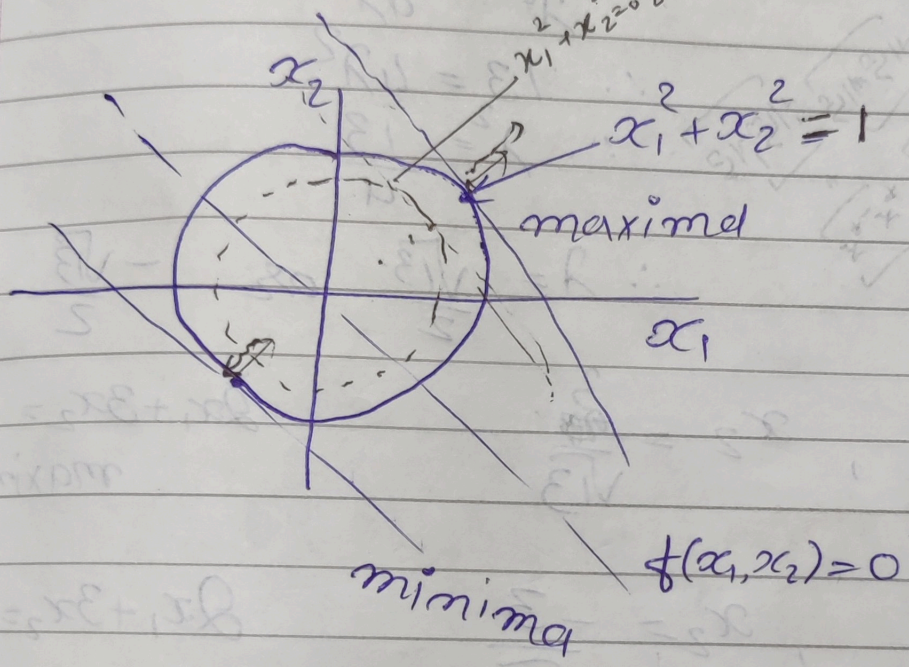
→ This can be solved using lagrange multiplier approach.

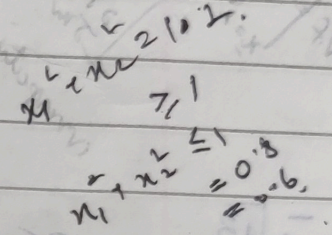
ex. Optimize $2x_1 + 3x_2$ such that $x_1^2 + x_2^2 = 1$

$f(x_1, x_2)$, objective function (linear)

constraint (non-linear)

$g(x_1, x_2) = x_1^2 + x_2^2 - 1$





→ Lagrangian $L(x_1, x_2, \lambda) = f(x_1, x_2) + \lambda\, g(x_1, x_2)$

lagrange multiplier

$= (2x_1 + 3x_2) + \lambda(x_1^2 + x_2^2 - 1)$

→ Differentiate w.r.t. x_1, x_2, λ and equate to zero and solve.

$\frac{\partial L(x_1, x_2, \lambda)}{\partial x_1} = 2 + 0 + 2x_1\lambda = 0$

$\therefore x_1 = -\frac{1}{\lambda}$

Gradient of g, i.e. $\nabla g\big|_{\underline{x}=\underline{x}^*}$ at $\underline{x}^* = \begin{bmatrix} \frac{-2}{\sqrt{13}} \\ \frac{-3}{\sqrt{13}} \end{bmatrix} = \begin{bmatrix} \frac{-4}{\sqrt{13}} \\ \frac{-6}{\sqrt{13}} \end{bmatrix}$

∇f and ∇g are in the same direction if for min gradients are in opposite direction

$$\frac{\partial L(x_1, x_2, \lambda)}{\partial x_2} = 0 + 3 + 2x_2\lambda = 0$$

$$\therefore x_2 = \frac{-3}{2\lambda}$$

$$\frac{\partial L(x_1, x_2, \lambda)}{\partial \lambda} = 0 + 0 + x_1^2 + x_2^2 - 1 = 0$$

$$\therefore x_1^2 + x_2^2 = 1$$

$$\therefore \frac{1}{\lambda^2} + \frac{9}{4\lambda^2} = 1$$

$$\therefore 13 = 4\lambda^2$$

$$\therefore \lambda^2 = \frac{13}{4}$$

$$\therefore \lambda = \frac{\sqrt{13}}{2} \text{ or } \frac{-\sqrt{13}}{2}$$

Gradient vector of f i.e. $\nabla f \Big|_{\underline{x}=\underline{x}^*} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

irrespective of $\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} 2/\sqrt{13} \\ 3/\sqrt{13} \end{bmatrix}$ or $\begin{bmatrix} -2/\sqrt{13} \\ -3/\sqrt{13} \end{bmatrix}$

$\lambda = \frac{-\sqrt{13}}{2}$ $\therefore x_1 = \frac{2}{\sqrt{13}}$, $x_2 = \frac{3}{\sqrt{13}}$

$2x_1 + 3x_2 = 3.6$ maxima

$\lambda = +\frac{\sqrt{13}}{2}$ $x_1 = \frac{-2}{\sqrt{13}}$, $x_2 = \frac{-3}{\sqrt{13}}$

$2x_1 + 3x_2 = -3.6$ maxima

+ve for minimum where we consider $\nabla f = -\lambda \nabla g$

because we have considered $f(x_1, x_2) + \lambda g(x_1, x_2) = L$

If we consider $\nabla f = \lambda \nabla g$ Then λ is -ve for minimum when we consider $f(x_1, x_2) - \lambda g(x_1, x_2) = L$

$\lambda > 0 \Rightarrow$ maxima and vice versa

* Consider the following problem which we call as primal optimization problem.

$$\min_{\underline{\theta}} f(\underline{\theta}) \quad \text{s.t. } g_i(\underline{\theta}) \le 0 \quad i = 1 \text{ to } k$$
$$h_i(\underline{\theta}) = 0 \quad i = 1 \text{ to } l$$

$\underline{\theta}$ called primal variables

→ To solve this problem, define generalized lagrangian.

$$L(\underline{\theta}, \underline{\alpha}, \underline{\beta}) = f(\underline{\theta}) + \sum_{i=1}^{k} \alpha_i g_i(\underline{\theta}) + \sum_{i=1}^{l} \beta_i h_i(\underline{\theta})$$

- α_i, β_i are called lagrange multiplier.

→ Consider the quantity

$$\theta_p(\underline{\theta}) = \max_{\underline{\alpha}, \underline{\beta}, \alpha_i \geq 0} L(\underline{\theta}, \underline{\alpha}, \underline{\beta})$$

↓ Primal

→ let some $\underline{\theta}$ be given if $\underline{\theta}$ violates any of the primal constraints. i.e. $g(\underline{\theta}) > 0$, $h_i(\underline{\theta}) \neq 0$ then $\theta_p(\underline{\theta}) = \infty$

- If primal constraints are satisfied

$\theta_p(\underline{\theta}) = f(\underline{\theta})$ → ~~It~~ $\theta_p(\underline{\theta})$ ~~selection~~ will be that $f(\underline{\theta})$ value which satisfies constraint. [See Example $x_1^2 + x_2^2 \geq 1$ ≤ 1

→ So we can write original primal problem as

$$\min_{\underline{\theta}, \underline{\alpha}} \theta_p(\underline{\theta})$$

maximize such that the condition are satisfied.

Find the max of L by choosing $\underline{\alpha}, \underline{\beta}$ with the constraints satisfied.

$$\therefore \min_{\underline{\theta}} \left(\max_{\underline{\alpha}, \underline{\beta}, \alpha_i \geq 0} L(\underline{\theta}, \underline{\alpha}, \underline{\beta}) \right)$$ ~~for convex f & g~~

↓ choose $\underline{\theta}$ max over α, β then $f(\theta)$ complete do this for any $\underline{\theta}$ — (i)

This will give optimum values of parameters $\theta^*, \alpha^*, \beta^*$

→ The dual problem can be written as

$$\max_{\underline{\alpha},\underline{\beta},\alpha_i \geq 0} \left[\min_{\underline{\theta}} L(\underline{\theta}, \underline{\alpha}, \underline{\beta}) \right] \quad \text{—(ii)}$$

dual variables

→ let at $\underline{\theta}^*, \underline{\alpha}^*, \underline{\beta}^*$ value of eq. (i) be p^* and value of eq. (ii) be d^*.

for some →

		for some		
$\underline{\theta}_1$	max = 25	$\underline{\alpha}_1$	$\underline{\beta}_1$	min = 2
$\underline{\theta}_2$	max = 26	$\underline{\alpha}_2$	$\underline{\beta}_2$	min = 3
$\underline{\theta}_3$	max = 27	$\underline{\alpha}_3$	$\underline{\beta}_3$	min = 4 →
$\underline{\theta}_4$	max = 88	$\underline{\alpha}_4$	$\underline{\beta}_4$	min = 1

for minimum of maximum

$p^* = \min(25, 26, 27, 88) = 25$

$d^* = \max(2, 3, 4, 1) = 4$

maximum of minimum

$d^* \leq p^*$

→ In certain conditions $d^* = p^*$.

What are the conditions under which $d^* = p^*$, so that one can get $\underline{\theta}$ either by using primal or dual formulation? of the problem. These conditions are called as Karush-Kuhn-Tucker (KKT) conditions listed below.

(1) $$\frac{\partial}{\partial \theta_i} \left[L(\underline{\theta}, \underline{\alpha}, \underline{\beta}) \right] \Big|_{\underline{\theta}=\underline{\theta}^*, \underline{\alpha}=\underline{\alpha}^*, \underline{\beta}=\underline{\beta}^*} = 0$$

all θ's

$i = 1$ to n

Note that I am using θ & $\underline{\theta}$ both for vector notation & parameters → Abuse of notation

② $\frac{\partial}{\partial \beta_i}[L(\underline{\theta}, \underline{\alpha}, \underline{\beta})]\Big|_{\underline{\theta}=\underline{\theta}^*, \underline{\beta}=\underline{\beta}^*, \underline{\alpha}=\underline{\alpha}^*} = 0 \quad i = 1 \text{ to } l$

③ $\alpha_i^* g_i(\underline{\theta}^*) = 0 \quad i = 1 \text{ to } k$

④ $g_i(\underline{\theta}^*) \leq 0 \quad i = 1 \text{ to } k$ (>0 for maxima)

⑤ $\alpha_i^* \geq 0 \quad i = 1 \text{ to } k$

→ These conditions will be satisfied if f & g are convex and all (f, g & h) are affine. (h is affine.)

(all linear (affine) fns are also convex)

* convex :- $f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$, $\alpha \in [0,1]$

f : a real fn f defined on interval is convex if for any two points x_1 & x_2

— $f(x) = x^2$ is a convex

affine :- $h(x) = ax + b$ is both convex and concave.

Let us come to SVM now

→ For SVM, we can write Lagrangian as

$$L(\underline{\theta}, \theta_0, \underline{\alpha}) = \frac{1}{2}\|\underline{\theta}\|^2 + \sum_{i=1}^{m} \alpha_i [1 - y^{(i)}(\underline{\theta}^T x^{(i)} + \theta_0)]$$

$f(\underline{\theta}, \theta_0)$ $\qquad$ $g_i(\underline{\theta}, \theta_0)$

→ Using the dual optimization approach

$$\max_{\underline{\alpha}} \left[\min_{\underline{\theta}, \theta_0} L(\underline{\theta}, \theta_0, \underline{\alpha}) \right]$$

, there are m number of α's α_i, $i = 1$ to m.

Our problem is $\min_{\theta, \theta_0} \frac{1}{2}\|\theta\|^2$ s.t $\{1 - y^{(i)}[\theta^T x^{(i)} + \theta_0)]\} \leq 0$

i.e. $\min_{\theta, \theta_0} f(\theta, \theta_0)$ s.t $g_i(\theta, \theta_0) \leq 0$, $i = 1$ to m → examples

NO $h(\theta)$ here

→ Consider minimization
- differentiate $L(\underline{\theta}, \theta_0, \underline{\alpha})$ w.r.t. each of θ_i, $i = 0$ to n and equate it to 0.

$$\frac{\partial}{\partial \theta_1}\left[L(\underline{\theta}, \theta_0, \underline{\alpha})\right] = 0$$

$$\vdots$$

$$\frac{\partial}{\partial \theta_n}\left[L(\underline{\theta}, \theta_0, \underline{\alpha})\right] = 0$$

$$\nabla_{\theta}\left(L(\underline{\theta}, \theta_0, \underline{\alpha})\right) = 0$$

θ to θ_n

$$\frac{\partial}{\partial \theta_0}\left[L(\underline{\theta}, \theta_0, \underline{\alpha})\right] = 0$$

$$\sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i y^{(i)} \underline{\theta}^T \underline{x}^{(i)}$$

Diff. w.r.t θ_0 gives

$$-\sum_{i=1}^{m} \alpha_i y^{(i)}$$

→ Hence we will get :

$$\underline{\theta} - \sum_{i=1}^{m} \alpha_i y^{(i)} \underline{x}^{(i)} = 0$$

vector

this is a vector

$$\begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$$

$$\theta_1 - \sum_{i=1}^{m} \alpha_i y^{(i)} x_1^{(i)} = 0$$ ← scalar

$$\vdots$$

$$\theta_n - \sum_{i=1}^{m} \alpha_i y^{(i)} x_n^{(i)} = 0$$

$$\boxed{\underline{\theta}^* = \sum_{i=1}^{m} \alpha_i y^{(i)} \underline{x}^{(i)}}$$

where α_i's are unknown

Before we go for getting α_i^*'s, $\alpha_i^* \geq 0$

→ differentiate w.o.r. θ_0

$$\therefore 0 = \sum_{i=1}^{m} \alpha_i y^{(i)}$$

→ By substituting $\underline{\theta}$, into $L(\underline{\theta}, \theta_0, \underline{\alpha})$

$$L(\underline{\theta}, \theta_0, \underline{\alpha}) = \frac{1}{2}\left[\sum_{i=1}^{m} \alpha_i y^{(i)} \underline{x}^{(i)}\right]^T \left[\sum_{j=1}^{m} \alpha_j y^{(j)} \underline{x}^{(j)}\right]$$

$$+ \sum_{i=1}^{m} \alpha_i \left[1 - y^{(i)}\left[\left(\sum_{j=1}^{m} \alpha_j y^{(j)} \underline{x}^{(j)}\right)^T \underline{x}^{(i)} + \theta_0\right]\right]$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \underline{x}^{(i)T} \underline{x}^{(j)}$$

(because $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$, $+0$)

→ This has to be maximized for α_i's with constraints as $\alpha_i \geq 0$ $i = 1$ to m, $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$

→ From lagrangian $L(\underline{\alpha}, \underline{\mu}, \underline{\lambda})$ objective function and can be solved. HW

→ Program is available for Quadratic optimization with linear constraints.

→ Once we get $\underline{\alpha}$

$$\underline{\theta}^* = \sum_{i=1}^{m} \alpha_i^* y^{(i)} \underline{x}^{(i)}$$

* looking at KKT conditions

- $\alpha_i^* g_i(\underline{\theta}^*, \theta_0^*) = 0 \quad i = 1 \text{ to } m$

- $g_i(\underline{\theta}^*, \theta_0^*) = 1 - y^{(i)}[\underline{\theta}^{*T} \underline{x}^{(i)} + \theta_0^*] \leq 0$

$\therefore y^{(i)}(\underline{\theta}^{*T} \underline{x}^{(i)} + \theta_0^*) \geq 1$

here $\underline{\theta} = (\theta_1, \ldots \theta_n)$

- $\alpha_i^* \geq 0$

→ If $\alpha_i > 0$ $\quad g_i(\underline{\theta}^*, \theta_0^*) = 0$

$\therefore y^{(i)}[\underline{\theta}^{*T} \underline{x}^{(i)} + \theta_0^*] = 1$

monzone

Lagrange multipliers are nonzero only for support vectors. i.e., data

support vectors

→ If $\alpha_i = 0$ then $g_i(\underline{\theta}^*, \theta_0^*) < 0$, non support vectors. (i.e., $y^{(i)}[\underline{\theta}^{*T} \underline{x}^{(i)} + \theta_0^*] > 1$)

This implies that → In getting only support vectors are contributing.

$\underline{\theta}^* = [\theta_1^*, \theta_2^* \ldots \theta_n^*]$ which means if we knows in advance which data points constitute Support vectors

α_i (nonzero) can be computed only for those data points

It is not possible to know which data point constitute support vectors

* For positive support vector:-

$\underline{\theta}^{*T} \underline{x}^{(i)} + \theta_0^* = 1$

$\therefore \theta_0^* = 1 - \underline{\theta}^{*T} \underline{x}^{(i)}$

- For negative support vector:-

$\theta_0^* = -1 - \underline{\theta}^{*T} \underline{x}^{(i)}$

- In general.-

$$\boxed{\theta_0^* = y^{(i)} - \underline{\theta}^{*T} \underline{x}^{(i)}}$$

→ This is a convex quadratic optimization problem so unique solution for $\underline{\theta}, \theta_0$

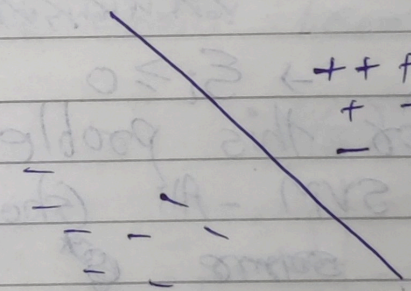
(An Advantage with SVM) → unlike neural nets.

* Few points about Lagragian that we got after substituting $\underline{\theta}$ ..
- Always $\underline{x}^{(i)}$ and $\underline{x}^{(i)^T}$ appeared as dot product.
- Unknowns are m which is irrespective of size of $\underline{x}$ (data vector) ↳ α_i's

What we did:

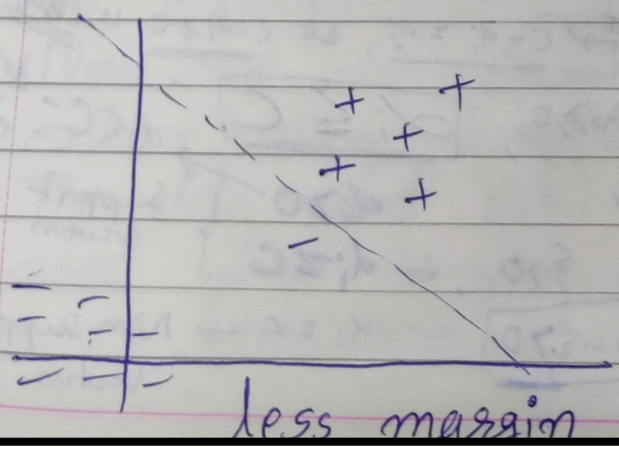
$$\min_{\underline{\theta}, \theta_0} \frac{1}{2} \|\underline{\theta}\|^2 \quad \text{s.t.} \quad y^{(i)} (\underline{\theta}^T \underline{x}^{(i)} + \theta_0) \geq 1 \quad i = 1 \text{ to } m$$

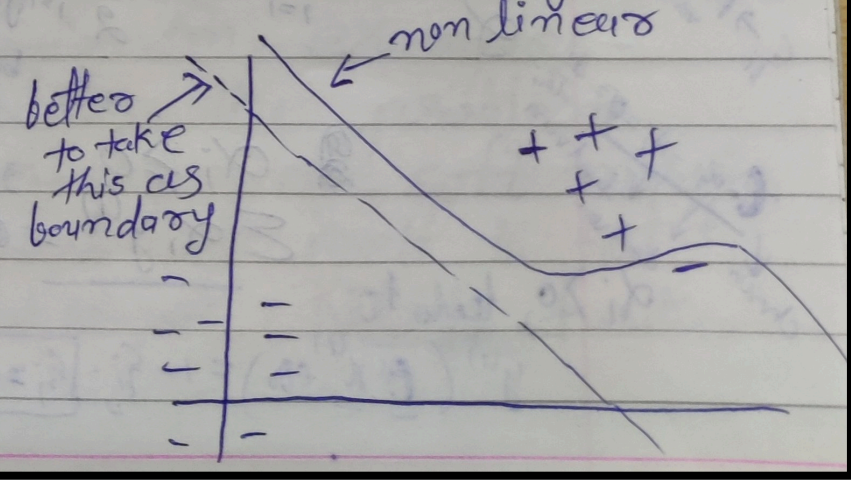
- This is called as hard margin classifier.
- If the data is not separable by hyperplane then this method has no solution.



→ To solve such a problem we use soft margin classifier.

✡ Soft margin SVM (SVM with regularization):





* Problem formulation :-
- We still get hyperplane taking care of outliers & non-linear data.
  → data that is not separable by linear boundary (hyperplane is linear boundary)

* objective function

$$\min_{\theta, \xi_i} \frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{m}\xi_i$$

→ If we don't include this we have no constraint since any point can violate the margin

Any point which violates the margin it adds as penalty in the objective fn & we try to minimize this objective fn.

Few data points are going to violate the strict constraint ≥ 1 so include ξ_i

C → regularization parameter

such that → $y^{(i)}(\theta^T x^{(i)} + \theta_0) \geq 1 - \xi_i$

ξ_i accounts for not held tightly in position → slack param

$\rightarrow \xi_i \geq 0$

H.W. lagrangian for this problem, dual form
- Soft margin SVM - Ali Ghodsi
→ This will give

$$\theta = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

In soft margin $0 \leq \alpha \leq C$

Convex Hence global minimum

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)}$$

ξ_i → primal variable

$\alpha_i - \lambda_i = 0 = \frac{dL}{d\xi_i}$ dual var, dual var, chosen

but → $C - \alpha_i = \lambda_i$, i.e. $C - \alpha_i \geq 0$ i.e. $C \geq \alpha_i$

$\alpha_i \geq 0$, $\lambda_i \geq 0$, $\boxed{\alpha_i \leq C}$

$\sum \alpha_i y^{(i)} = 0$

$\alpha_i > 0$, leads to $y^{(i)}(\theta^T x^{(i)} + \theta_0) = 1 - \xi_i$, $\boxed{\xi_i = 0 \text{ or } \xi_i > 0}$

$\alpha_i < C$ → do not violate the margin

$\alpha_i > 0$ → support vectors

$\xi_i > 0$ ← $\alpha_i = C$

$\alpha_i = 0$ → Non support

→ data is highly non-linear, then use SVM with Kernel function. for non-linearly separable data points. Kernel trick works here.

* Kernel Trick :-
- Hard margin classifier – linearly separable
- Soft margin classifier = slack variable ξ_i are used
  - takes care of non linear separable but still not the best classifier
- Use of Kernel function – Best SVM classifier with high accuracy.

→ When the data is not linearly separable, project it onto higher dimensional space where the data becomes linearly separable.

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \phi(\underline{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix}$$

Ex → We saw that using polynomial regression where we used 3 features x_1, x_2, x_1x_2 gives us better results

lower dimensional data → higher dimensional data

i.e. lower to higher dimensional mapping

m examples → m examples

→ Now, we can formulate the problem either using hard margin or soft margin
- The data used $\phi(\underline{x})$. But ϕ itself is not used ~~due to the use of~~ KERNEL function.
  What we use is