Support Vector Machines (see Lagrangian, primal & dual from notebook)

The Lagrangian → for constrained optimiz^n problem

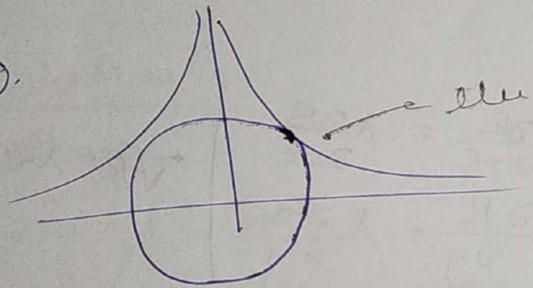
Optimiz^n problems → constrained, unconstrained

e.g. $\underset{max}{f(x,y)} = x^2 e^y y = c$ } unconstrained

e.g. $\underset{max}{f(x,y)} = x^2 e^y y = c$ } constrained optimiz^n problem

$g(x,y) = x^2 + y^2 = 4$ → constraint.

e.g.

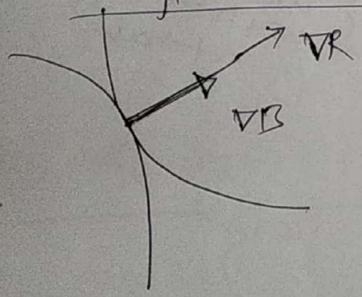


the place where line of f is perfectly parallel to this contour of g.

generally $f = R$ (revenue) → we want to maximize revenue

$g = B$ (budget) → you have some ~~sort of~~ dollar limit by Budget.

different approach to solve this problem



the two gradients must be in the same dir^n or proportional.

i.e. $\nabla R = \lambda \nabla B$ ⟨at optimal cond^n⟩

so basically Eq^n needed

$$\boxed{\begin{aligned} &\nabla R = \lambda \nabla B \\ \& \quad &B(x,y) = x^2 + y^2 = 4 \end{aligned}}$$

gives two eq^ns, one more

$4 \to b$

lagrangian is just a way to pack up these two eqns into a single entity

$$\boxed{L(x,y,\lambda) = R(x,y) - \lambda\,(B(x,y) - b)}$$

lagrangian multiplier optimizn funcn constraint.

Now $\nabla L = 0$ → gives (h^*, s^*, λ^*) or (x^*, y^*, λ^*)

$$\begin{bmatrix} \partial L/\partial x \\ \partial L/\partial y \\ \partial L/\partial \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial x} = \frac{\partial R}{\partial x} - \lambda \frac{\partial B}{\partial x} = 0 \qquad \frac{\partial R}{\partial x} = \lambda \frac{\partial B}{\partial x}$$

$$\frac{\partial L}{\partial y} = \frac{\partial R}{\partial y} - \lambda \frac{\partial B}{\partial y} = 0 \qquad \frac{\partial R}{\partial y} = \lambda \frac{\partial B}{\partial y} \qquad \nabla R = \lambda \nabla B$$

$$\frac{\partial L}{\partial \lambda} = B(x,y) - b = 0 \qquad B(x,y) = b.$$

Primal standard form

$$\max Z = \sum_{j=1}^{n} c_j x_j$$

s.t. $\sum_{j=1}^{n} a_{i,j}\, x_j \le b_i \quad i = 1 \text{ to } m$ } main constrain is less than

$x_j \ge 0, \quad j = 1 \text{ to } n$

Dual standard form

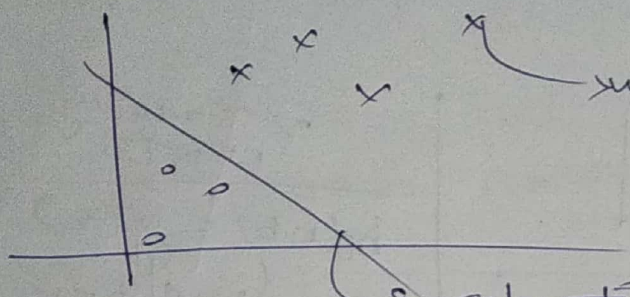
$$\min \; W = \sum_{i=1}^{m} u_i b_i$$

st. $\sum_{i=1}^{m} a_{j,i}\, u_i \ge c_j, \quad j = 1 \text{ to } n$ } both greater than

$u_i \ge 0 \quad\quad i = 1 \text{ to } m$

Support Vector Machines (classifier)

We want correct and confident prediction



most confident

separating Hyperplane.

Hyperplane

The subspace whose dimension is one less than its ambient space.

Notation

hypothesis

$h_{w,b}(x) = g(w^T x + b)$

$y \in \{-1, 1\}$

instead of $\{0, 1\}$

using perceptron algorithm

we can get directly

$g(w^T x + b) = 1$ or -1

i.e. $\left. \begin{array}{ll} g(z) = 1 & \text{if } z \geq 0 \\ g(z) = -1 & \text{o/w} \end{array} \right\}$

Earlier

$h_\theta(x) = g(\theta^T x)$

$h_\theta(x) = P(y=1 \mid x; w, b)$

$\left. \begin{array}{l} h_\theta(x) \geq 0.5,\ \theta^T x \geq 0 \\ \text{or } \theta^T x \gg 0 \end{array} \right\} y = 1$

confident predict

Functional Margin

- Since we want confident prediction, we define functional margin as -

for 1 TE

$$\boxed{\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)}$$

for all TE

$$\hat{\gamma} = \min_{i=1 \text{ to } m} \hat{\gamma}^{(i)}$$

z is actually $w^T x + b$

$h_\theta(w,b) = g(z)$

(but here ±1 perceptron algorithm)

{ If $y^{(i)} = 1$, then $w^T x + b$ must be +ve large for confident prediction. (z must be large)

if $y^{(i)} = -1$ then $w^T x + b$ -ve large.

Not true

- functional margin can only be used to check correct prediction or not but not confidence

i.e. if

$$\boxed{\begin{array}{ll} \hat{\gamma}^{(i)} > 0 & \text{correct prediction} \\ \hat{\gamma}^{(i)} < 0 & \text{incorrect prediction} \end{array}}$$

because:

$w^T x + b$ and $2w^T x + 2b$ changes the value of functional margin but does not change anything geometrically.

i.e. $g(w^T x + b) = g(2w^T x + 2b)$

for training set $S = \{(x^{(i)}, y^{(i)}) ; i = 1 \text{ to } m\}$

we want to find

$$\boxed{\hat{\gamma} = \min_{i=1 \text{ to } m} \hat{\gamma}^{(i)}}$$

* Among all TE, what is the min^m value of FM is $\hat{\gamma}$ of complete Training set.

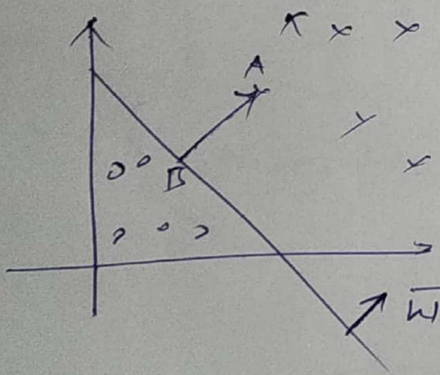
Geometric Margin

(can be actually used to measure confidence wrt a given line

bcz remains const. by changing just coeff.

→ we are measuring confident classific^n wrt a given line.

i.e. we are trying to find a line/classifier with max^m confidence or GM.



for one TE

$$\boxed{\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|} \right)}$$

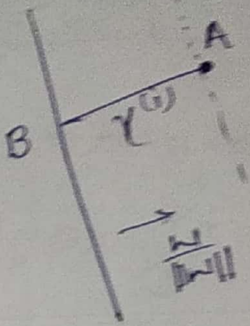
for all TE

$$\boxed{\gamma = \min_{i=1 \text{ to } m} \gamma^{(i)}}$$

Derivatⁿ of GM :- decision boundary

Let point A is $x^{(i)}$

Unit vector ⊥r to DB $= \frac{w}{\|w\|}$



Vector AB $= \gamma^{(i)} \frac{w}{\|w\|}$

so B $= x^{(i)} - \gamma^{(i)} \frac{w}{\|w\|}$

B lies on DB

so, satisfy $w^T x + b = 0$

$$\Rightarrow w^T\left(x^{(i)} - \gamma^{(i)} \frac{w}{\|w\|}\right) + b = 0$$

$$\Rightarrow \boxed{\gamma^{(i)} = \frac{w^T x^{(i)} + b}{\|w\|}} = \frac{\hat{\gamma}^{(i)}}{\|w\|}$$

Note

$GM(\gamma) \leq FM(\hat{\gamma})$

$\gamma^{(i)} \leq \hat{\gamma}^{(i)}$ i = 1 to m.

Optimiz^n problem (simplification)

(1) $\max_{\gamma, w, b} \ \gamma$

s.t. $y^{(i)}(w^T x^{(i)} + b) \geq \gamma$, $i = 1$ to m

$\|w\| = 1$

max of GM
GMTE ≥ GM
so $\|w\| = 1$

here the constraint $\|w\| = 1$ is nonconvex
(→ not possible to get (converge) global minima/optimum

(2) $\max_{\hat{\gamma}, w, b} \ \dfrac{\hat{\gamma}}{\|w\|}$

s.t. $y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}$ $i = 1$ to m

max of GM
FMTE ≥ FM

here the objective fucn $\dfrac{\hat{\gamma}}{\|w\|}$ is non-convex.

Note: since we can scale GM (numer.) any way we will choose $\hat{\gamma} = 1$

(3) $\max_{\gamma, w, b} \ \dfrac{1}{\|w\|}$

$y^{(i)}(w^T x^{(i)} + b) \geq 1$ $i = 1$ to m

Also

$$\min_{w,b} \ \frac{1}{2}\|w\|^2$$

s.t. $y^{(i)}(w^T x^{(i)} + b) \geq 1$

→ convex quadratic objective func^n
& linear constraint
its sol^n gives us optimal margin classifier
→ Not notin primal form

Optimiz^n problem using lagrange's dual form

Benefits -

i) Used to show svm depends on very small no. of support vectors

ii) The optimiz^n problem can be expressed in terms of inner product only:

$\langle x^{(i)}, x^{(j)} \rangle$ or $\left((x^{(i)})^T x^{(j)} \right)$

{ for Kernel trick, no need to find ϕ and going high dimension }

lagrange duality → solving constrained optimiz^n problem

$$\min_w f(w)$$

s.t. $h_i(w) = 0$, $i = 1, \dots, l$

solving it using lagrange multiplier method,

$$L(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

→ lagrangian multiplier

we would then find PD of L to zero.

$$\boxed{\frac{\partial L}{\partial w_i} = 0, \quad \frac{\partial L}{\partial \beta_i} = 0}$$

and solve for w and β

Generalizing primal optimiz^n problem to dual optimiz^n problem -

primal.

$$\min_{w} f(w)$$

s.t. $g_i(w) \le 0$, $i = 1$ to k
$h_i(w) = 0$, $i = 1$ to l } primal constraints

Generalized Lagrangian

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

Lagrangian multiplier.

Max के साथ lagrange multiplier की constraint

consider, θ_p where p stands for primal,

$$\theta_p(w) = \max_{\alpha, \beta : \alpha_i \ge 0} L(w, \alpha, \beta)$$

$$\theta_p(w) = \max_{\alpha, \beta : \alpha_i \ge 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\theta_p(w) = \begin{cases} f(w) & , \text{if } w \text{ satisfies primal constraints} \\ \infty , & \text{otherwise} \end{cases}$$

If w satisfies primal constraints —

$$\boxed{\min_{w} f(w) = \min_{w} \theta_p(w) = \min_{w} \max_{\alpha, \beta : \alpha_i \ge 0} L(w, \alpha, \beta)}$$

let $\boxed{p^* = \min_{w} \theta_p(w)}$ → value of the primal problem.

Dual optimiz^n problem (its transpose)

$$\theta_D(\alpha, \beta) = \min_{w} L(w, \alpha, \beta)$$

$$\max_{\alpha, \beta; \alpha_i \geq 0} \theta_D(\alpha, \beta) = \max_{\alpha, \beta; \alpha_i \geq 0} \min_{w} L(w, \alpha, \beta)$$

Order of minmax changes only in dual = max min multipliers fun

let $d^* = \max_{\alpha, \beta; \alpha_i \geq 0} \min_{w} L(\alpha, \beta, w) = \max_{\alpha, \beta; \alpha_i \geq 0} \theta_D(\alpha, \beta)$

$\therefore$ maxmin $\leq$ minmax of a fun

$$d^* \leq p^*$$

$$\max_{\alpha, \beta; \alpha_i \geq 0} \min_{w} L(\alpha, \beta, w) \leq \min_{w} \max_{\alpha, \beta; \alpha_i \geq 0} L(w, \alpha, \beta)$$

So, there must be one soln for equality cond^n w^*, α^*, β^*

s.t. $d^* = p^* = L(w^*, \alpha^*, \beta^*)$

at Equality cond^n it must satisfy KKT (Karush-Kuhn-Tucker) conditions $\rightarrow$

KKT (Karush-Kuhn-Tucker conditions)

$$\frac{\partial}{\partial w_i} L(w^*, \alpha^*, \beta^*) = 0 \qquad i = 1 \text{ to } n$$

$$\frac{\partial}{\partial \beta_i} L(w^*, \alpha^*, \beta^*) = 0 \qquad i = 1 \text{ to } L$$

$$\left.\begin{aligned} \alpha_i g_i(w^*) &= 0 \\ \alpha_i &\geq 0 \end{aligned}\right\} \; i = 1 \text{ to } K$$

$\alpha_i g_i(w^*) = 0 \rightsquigarrow$ ★★★ (used for support vectors)

$$g_i(w^*) \leq 0$$

$w \to n \qquad \alpha \to K \qquad \beta \to L$

Optimal Margin Classifier

Note:-
$\alpha_i > 0$ only if
$g_i(w^*) = 0$

Primal form

Optimiz^n problem standard primal form —

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

s.t $-y^{(i)}(w^T x^{(i)} + b) + 1 \le 0 \rightarrow g_i(w)$

Dual form

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i \left[1 - y^{(i)}(w^T x^{(i)} + b)\right]$$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1\right]$$

$\because \alpha_i > 0$ if $g_i(w^*) = 0$

— so only the one's for which functional margin is 1 $\alpha_i = 1$

— for others $\alpha_i = 0$

— support vectors

(less calculat^n does not consider other TEs for DB calcul^n)

for showing support vectors needed just transform into dual form and show KKT cond^n $\alpha_i g_i(w) = 0$

Representation of lagrangian only in terms of inner product (Kernel trick) (II)

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right] \quad \text{--- (1)}$$

Dual = maxmin

so, for min w.r.t. w, b take gradient

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

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

so,

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

$$\boxed{w^T = \sum_{i=1}^{m} \alpha_i y^{(i)} (x^{(i)})^T}$$

and

$$\frac{\partial L(w,b,\alpha)}{\partial b} = -\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Substitute in (1) to get the form

$$L(w,b,\alpha) = \frac{1}{2} \sum_{\substack{i=1 \\ j=1}}^{i,j=m} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)} (x^{(j)})^T + \sum_{i=1}^{m} \alpha_i$$

$$- \sum_{\substack{i=1 \\ j=1}}^{i,j=m} \alpha_i y^{(i)} x^{(i)} \alpha_j y^{(j)} (x^{(j)})^T$$

$$- b \cancel{\sum_{i=1}^{m} \alpha_i y^{(i)}}^{\,0}$$

$$\boxed{L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)} (x^{(j)})^T}$$

$$\boxed{L(w,b,\alpha) = \max_{\alpha} \theta_D(w,b) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle}$$

$$\alpha_i \geq 0, \quad i = 1 \text{ to } n$$

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

Only expressed in terms of inner product, so, when replaced to $\phi(x^{(i)})$, $\phi(x^{(j)})$ we can directly use $K(x^{(i)}, x^{(j)}) = \phi(x^{(i)}) \phi(x^{(j)})$ ...

For a new example → we will predict 1 if $w^T x + b > 0$

$$w^T x + b = \left(\sum_{i=1}^{m} \alpha_i x^{(i)} y^{(i)}\right)^T x + b$$
$$= \sum_{i=1}^{m} \alpha_i y^{(i)} (x^{(i)})^T x + b$$
$$= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b$$

↳ again only inner product
can use kernel directly.

$$\boxed{b^* = -\frac{\max_{i: y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i: y^{(i)}=1} w^{*T} x^{(i)}}{2}}$$

Kernel Trick

we can use many a time feature maps $\phi(x)$ from attributes to features

e.g. x to say x^2

In that case inner product in 4 & 5 will be replaced by $\langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$.

Note: we define kernel
$$K(x, z) = \phi(x)^T \phi(z)$$
$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

↳ effective way to calculate $K(x, z)$ can get SVMs to learn in the high dimension feature space given by ϕ without even having to explicitly find or represent vector $\phi(x)$.

Types of Kernels

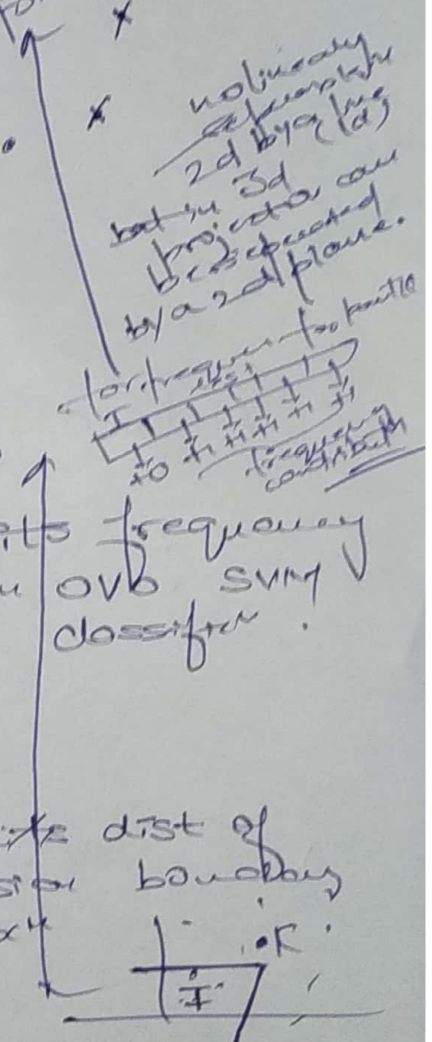
Use linear Kernel if it is linearly seperable or features is very large as otherwise go in high dimension until linearly seperable and computationally very expensive can overfit also

funda → 3 classes for linear better.

i) if $m > n$ linear if linearly seperable try others too.

ii) if $n >> m$ use linear (try to only)

linear.
RBF
Gaussian
tanh.



Majority voting eg in the positive points the point occurs most is the label of the point/TE

⇑

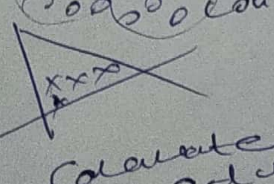
One v/s one SVM — say 6 categories

finds the or classifies in category to which its frequency is maximum in OVO SVM classifier.

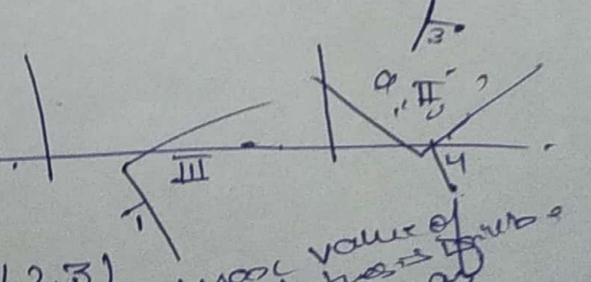
$\frac{n(n-1)}{2}$ classifiers

One v/s all SVM

classify to a category it to the class in which ~~with the~~ dist of decision boundary is max^n



contradiction why I & III
∴ dist 3 > 1
so set it in III class.



Calculate n classifiers and get the max one for $h_\theta^{(i)}(x) = P(y = i \mid x; \theta)$ $(i = 1, 2, 3)$ max value of hypothesis to make one at

Error Metrics → F_1 score is used → for misclassification error.

$$\text{Precision}_j = \frac{TP}{TP + FP} \quad \text{for class } j$$

positive

$$\text{recall}_j = \frac{TP}{TP + FN} \quad \text{for class } j$$

$$\boxed{F_{1j} = 2 \times \frac{\text{Precision}_j \times \text{recall}_j}{\text{Precision}_j + \text{recall}_j}}$$

We got 96% on test data and 99% on training data.

Cost function.

$$J = \frac{1}{m} \sum_{i=1}^{m} \max\left(0,\ 1 - y^{(i)}\left(w^T x^{(i)} + b\right)\right) + \lambda \|\vec{w}\|^2$$

↓ regularizer

* for one vs all for new e.g. answer i/p x to make predict pick i that maximizes

$$\max_i h_\theta^{(i)}(x)$$