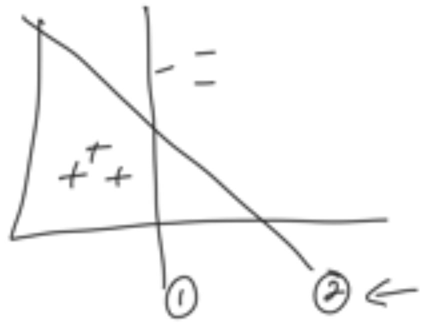
Last time: max margin



← because it's farther from the data. generalizes in the face of perturbations

$\text{margin}(x_n, y_n) = \dfrac{y_n(w^T x_n + w_0)}{\|w\|_2}$ } relative distance } absolute distance

two formulations:

hard-margin (separable)

$$\boxed{\min_{w, w_0} \tfrac{1}{2}\|w\|_2^2 \quad \text{s.t. } y_n(w^T x_n + w_0) \geq 1}$$

soft margin

$$\min_{w, w_0} \tfrac{1}{2}\|w\|_2^2 + c \sum \xi_n$$
$$\text{s.t. } y_n(w^T x_n + w_0) \geq 1 - \xi_n, \quad \xi_n \geq 0$$

↳ both convex! in high dimensions, solving these can still be slow

today: convert these into "dual" formulation [taste of optimization]

↳ incorporating kernels/bases

apply Lagrangian:

$$\min_{w, w_0} \max_{\alpha, \alpha \geq 0} \boxed{\tfrac{1}{2}\|w\|_2^2} - \sum_n \boxed{\alpha_n} \boxed{\left[y_n(w^T x_n + w_0) - 1\right]}$$

thought experiment: consider new obj

old obj + α II (constraint met)

⇕

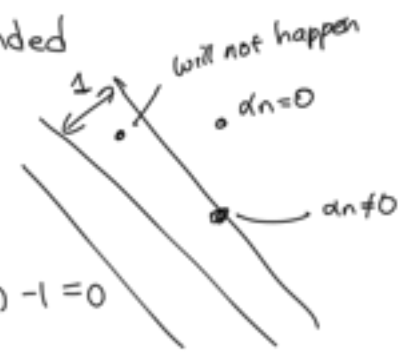
old obj + α amount of violation.

∇old $\pm \nabla$ violation $= 0$.

1) if constraint is violated, $y_n(w^T x_n + w_0) - 1 < 0$
then α can become very large & make obj unbounded

2) if the constraint is met & has some slack,
$y_n(w^T x_n + w_0) - 1 > 0$ (NOT equal to 0)
$\alpha_n = 0$

only case where $\alpha_n \neq 0$ will be when $y_n(w^T x_n + w_0) - 1 = 0$



Now notice

3) given (2), at optimality, we recover the original objective

Next steps: by property known as strong duality, we can swap the max & min

$$\boxed{\max_{\alpha} \min_{w, w_0} \overbrace{\tfrac{1}{2}\|w\|_2^2 - \sum_n \alpha_n \left[y_n(w^T x_n + w_0) - 1\right]}^{\mathcal{L}}}$$

goal: 1st let's solve the min. problem, next deal w/ the max

- $\frac{\partial L}{\partial w} = w - \sum_n \alpha_n y_n x_n = 0 \Rightarrow \boxed{w = \sum_n \alpha_n y_n x_n}$
- $\frac{\partial L}{\partial w_0} = -\sum_n \alpha_n y_n = 0$ } new constraint

$$\max_\alpha \underbrace{\frac{1}{2}\left(\sum_n \alpha_n y_n x_n\right)^T\left(\sum_{n'} \alpha_{n'} y_{n'} x_{n'}\right)}_{w^Tw} - \underbrace{\sum_n \alpha_n y_n \left(\sum_{n'} \alpha_{n'} y_{n'} x_{n'}\right)^T x_n}$$ } sub. for w

$$- \underbrace{\sum_n \alpha_n y_n w_0}_{w_0 \sum_n \alpha_n y_n \to 0} + \sum_n \alpha_n$$

$$\boxed{\max_\alpha -\frac{1}{2}\sum_{n,n'} \alpha_n \alpha_{n'} y_n y_{n'} \boxed{x_{n'}^T x_n} + \sum_n \alpha_n \quad \text{s.t.} \quad \sum_n \alpha_n y_n = 0, \ \alpha_n \geq 0}$$

new optimization problem in terms of α (instead of w!)
easy to solve! (convex)

Note: soft margin: constraint $\alpha_n \geq 0$ becomes $C \geq \alpha_n \geq 0$

once we solve for α^*, we can recover w, w_0:

$$\boxed{w^* = \sum_n \alpha_n^* y_n x_n}$$

↳ implication: $w^T x + w_0$

$$\Rightarrow \left[\underbrace{\sum_n \alpha_n^*}_{\text{weight per datum in training set}} \underbrace{y_n}_{\text{vote of that pt.}} \underbrace{\boxed{x_n^T x}}_{\text{how similar are } x_n \text{ to test pt. } x} + w_0 \right.$$

typically # of $\alpha_n \geq 0$ is small... so sparse classifier / fast at test time.

finally w_0: take any n s.t. $\alpha_n > 0 \Rightarrow y_n(w^T x_n + w_0) = 1$
sub in ↑ solve for w_0

$$w_0 = \frac{1}{y_n} - w^{*T} x_n$$

D doesn't seem to be gone...
if we kernelize: $x \to \phi(x)$, we still need to deal w/ the dimensionality of ϕ

Notice! All computations ONLY require $x^T x'$ or $\phi(x)^T \phi(x')$

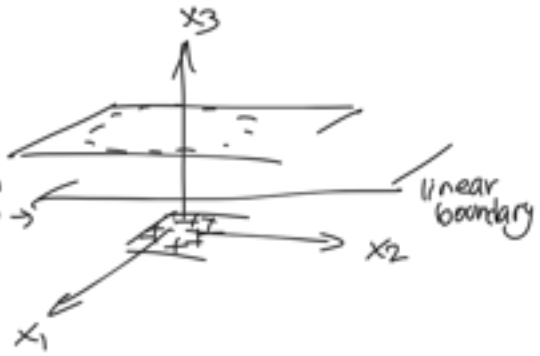
idea: we can define $\underbrace{K(x,x')}_{\text{kernel function}} = \phi(x)^T \phi(x')$

reminder: why bases help:

hard → soft

NOT linearly sep.

for ex, can move 2D → 3D



how to choose $\phi(x)$? expert, NN,

today: highly expressive spaces that correspond to our notions of similarity

Let's talk about $K(x,x')$:

ex: $K(x,x') = x^T x'$ (trivial)

ex. $K(x,x') \Rightarrow \quad \phi: [x_1^2 \; x_1x_2 \cdots x_1x_D \;\; x_2x_1 \; x_2^2 \cdots x_2x_D \cdots x_D^2]$

$K(x,x') = \sum_{d,d'} x_d x_{d'} x'_d x'_{d'}$

ex. $\boxed{K(x,x')} = \underline{(1+x^Tx')^q}$ Polynomial Kernel (corresponds to a ϕ that is exponential in q)

ex. $K(x,x') = \underline{\exp\left(\frac{-\|x-x'\|_2^2}{2\sigma^2}\right)}$ RBF (Radial basis function) corresponds to an ∞-D ϕ

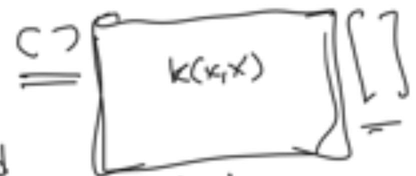
Can we just "create" some $K(x,x')$? "similarity function"

How do we know that it is valid?

Mercer's theorem:

Matrix form

- def matrix $\mathbb{K}$ be an $N \times N$ matrix where $\mathbb{K}_{nn'} = K(x_n, x_{n'})$ $\underline{\underline{[\;]}}$



- if K corresponds to a valid kernel, then $\mathbb{K}$ is positive semi-definite for any choice of input set $\{x_1 \ldots x_N\}$ $\underbrace{(z^T\mathbb{K}z \geq 0 \;\; \forall z)}$

$\Downarrow$

$$\int_{x,x'} f(x) K(x,x') f(x') \, dx dx' \geq 0 \quad \forall f \quad \text{(nice)}$$

then $K(x,x')$ is a valid kernel $\Leftrightarrow$ is an inner product in some space.

$\boxed{C}$ Thought exercise: consider RBF kernel w/ diff σ's

$$\exp\left(\frac{\|x-x'\|_2^2}{-2\sigma^2}\right)$$

as well as linear kernel $K(x,x') = x^Tx'$

overfit and how?

• can these overfit?

$$\sum a_n y_n \underbrace{K(x_n, x)} + w_0$$

Svm