CS 181 Max Margin Methods 2021

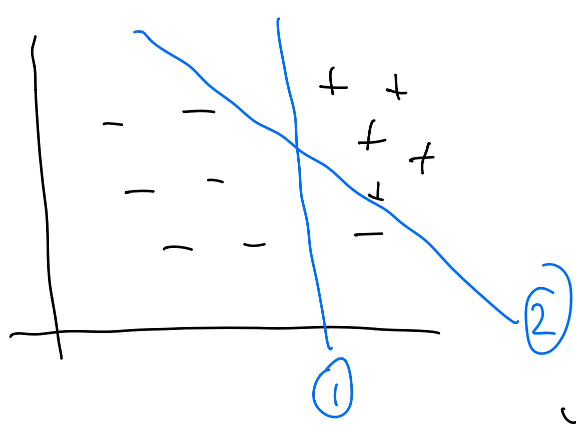
NNs	SVMs
representation learning	basis engineering
excellent perf	very good performance
non-convex / costly	convex / easy to train
hard to interpret	simple, interpretable
an "art"	coherent theory

<u>Setting</u> Binary classification

discr.

$$\hat{y} = \begin{cases} +1, & \text{if } \underline{w}^T \underline{x} + w_0 > 0 \\ -1 & \text{otherwise} \end{cases}$$

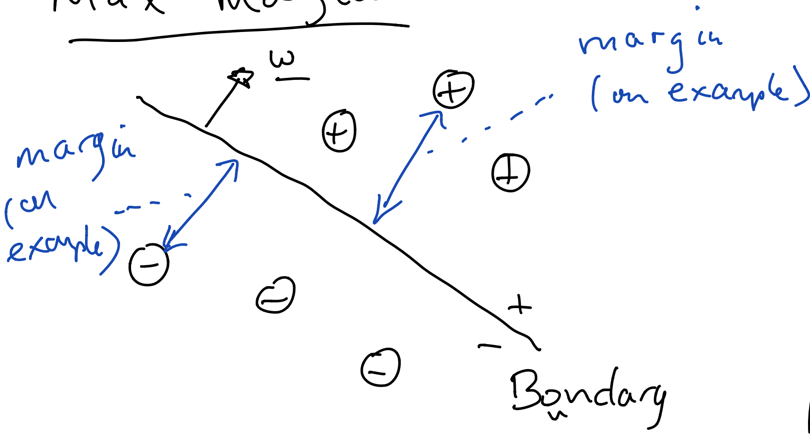
For <u>now</u> assume seperable data



Prefer 1 to 2

↳ generalize better (small perturbation to data will not matter)

Max margin

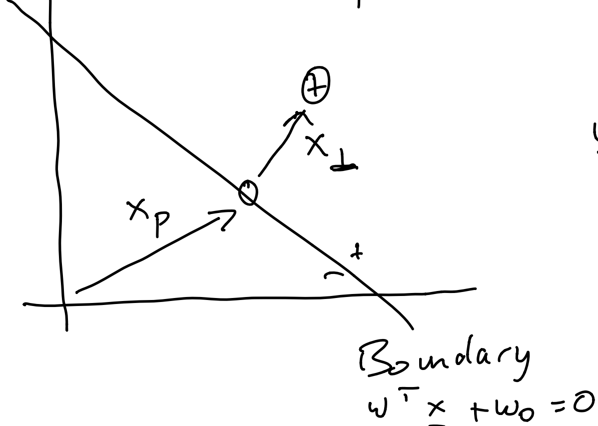


margin on correctly classified example is the absolute, normalized, orthogonal distance to boundary

"margin on data" = min margin on correct examples

Goal: find seperator that maximizes margin

Geometry



$$\underline{x} = \underline{x}_p + \underline{x}_\perp = \underline{x}_p + r \frac{\underline{w}}{\|\underline{w}\|_2}$$

$$\underline{w}^T \underline{x} = \underline{w}^T \underline{x}_p + r \frac{\underline{w}^T \underline{w}}{\|\underline{w}\|_2}$$

$$= -w_0 + r \|\underline{w}\|_2$$

$$\boxed{r = \frac{\underline{w}^T \underline{x} + w_0}{\|\underline{w}\|_2}}$$

Generally (pos ex. + neg ex.):

$$\text{margin}(x_n, y_n) = y_n \left(\frac{\underline{w}^T \underline{x}_n + w_0}{\|\underline{w}\|_2} \right) \quad (>0)$$

(*)

Note ① Margin is invariant to multiplying $(\underline{w}, w_0)$ by scalar $\beta > 0$

[But note that $y_n(\underline{w}^T x_n + w_0)$ increases!]

② (*) is negative if used on a misclassified example

$\boxed{\text{Hard max-margin formulation}}$

[1] $$\max_{\underline{w}, w_0} \left[\min_n \; y_n \left(\frac{\underline{w}^T \underline{x}_n + w_0}{\|\underline{w}\|_2} \right) \right]$$

→ will find a separator!

→ looks "ugly"

[2] By invariance to scaling by $\beta > 0$, w.l.o.g. to impose $y_n(\underline{w}^T \underline{x}_n + w_0) \geq 1$.

$$\max_{\underline{w}, w_0} \frac{1}{\|\underline{w}\|_2} \min_n y_n(\underline{w}^T \underline{x}_n + w_0)$$

$$\text{s.t. } y_n(\underline{w}^T \underline{x}_n + w_0) \geq 1 \quad \text{for all } n$$

<u>Equivalent</u> { one or more constraints will bind }

m

$$\max_{\underline{w},\, w_0} \frac{1}{\|\underline{w}\|_2}$$

s.t. $y_n(\underline{w}^T \underline{x}_n + w_0) \geq 1$ for all n

$\boxed{\text{Hard-margin formulation}}$

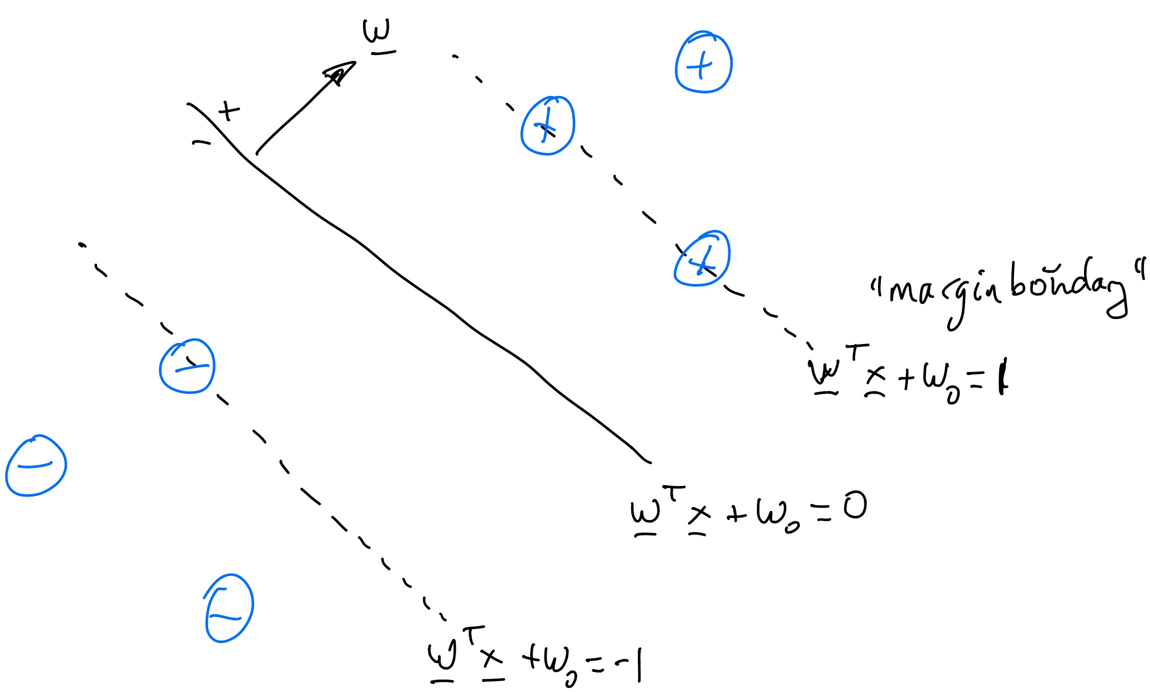
[Note, can also write $\min \frac{1}{2}\|\underline{w}\|_2^2$!]

$$\min_{\underline{w},\, w_0} \|\underline{w}\|_2^2$$

s.t. $y_n(\underline{w}^T \underline{x}_n + w_0) \geq 1$, all n

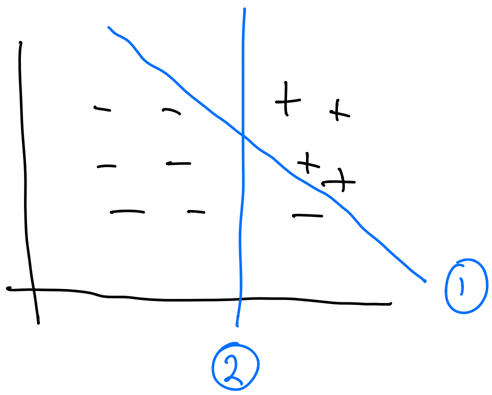
Nice! Convex (quadratic objective, linear constraints),

[margin $= \frac{1}{\|\underline{w}\|_2}$]



$\boxed{\text{Soft-margin formulation}}$

→ Regularization
→ Non-separable data

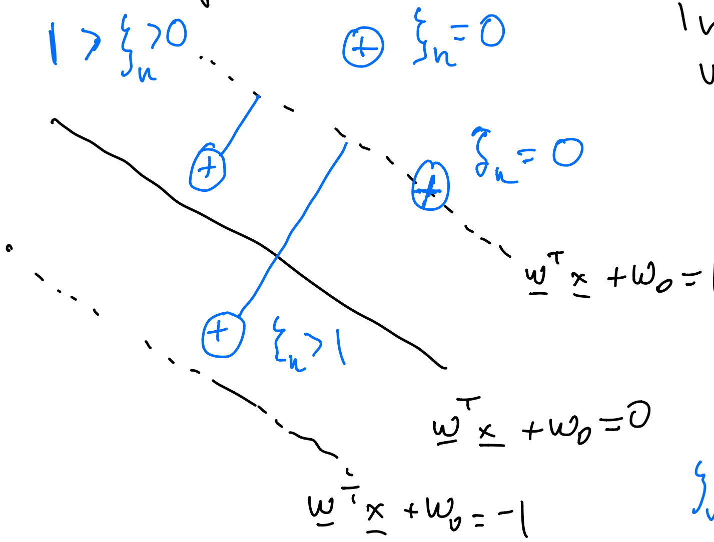


May prefer ② over ①, even though ② is not separating

(and data may not be separable!)

Relaxed formulation

(~ xi)



Introduce new variable $\xi_n \geq 0$, each n

"How much is $\underline{x}_n$ on the 'wrong side' of the margin boundary"

$\xi_n = 0$ correct class

$0 < \xi_n < 1$ correct class but smaller margin than $1/\|\underline{w}\|_2$

$1 < \xi_n$ incorrect class

Soft-margin formulation

For some $C > 0$ (regularization parameter)

$$\min_{\underline{w}, w_0, \underline{\xi}} \quad \frac{1}{2}\underline{w}^T\underline{w} + C\sum_n \xi_n$$

$$\text{s.t.} \quad y_n(\underline{w}^T\underline{x} + w_0) \geq 1 - \xi_n, \quad \text{all } n$$

$$\xi_n \geq 0, \quad \text{all } n$$

- "Pretend" margin $1/\|\underline{w}\|_2$ (ignoring examples live "in the margin")
- Allows misclassified points
- Large C ($C = \infty$ is hard margin)
  - → less regularization (try to get correct classification all all n)
- Smaller C
  - → Better (larger) $1/\|\underline{w}\|_2$ ("margin") more mistakes

Equivalently

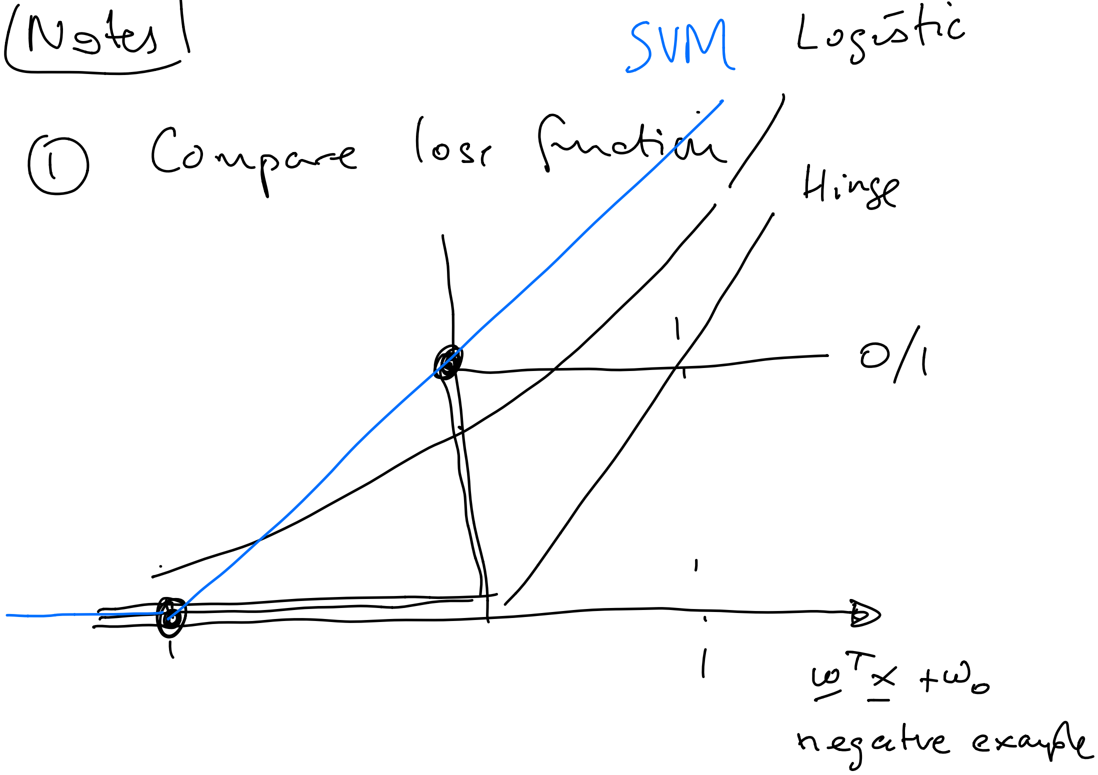
$$\min_{\underline{w}, w_0} \quad \frac{1}{2}\underline{w}^T\underline{w} + C\sum_n \max(0, 1 - y_n(\underline{w}^T\underline{x}_n + w_0))$$

Convex! $\cup$ +

+ diff (almost everywhere) $\Rightarrow$ SGD

Notes

1. Compare loss function



Why is it better than logistic?

[?] Quadratic + linear constraint

↳ Duality math; works with basis functions nicely

2. Large margins help to generalize
(think about small perturbation)