Regression & Classification

Losses: MSE, L1   0/1, hinge, logistic reg.

Models: linear models, basis models, neural networks

Idea: choose loss, model that make sense for the problem, but still computationally tractable.

Broader theme: we want models/losses that allow us to <u>generalize</u>.

Today: <u>loss</u> that (a) has some nice computational properties when paired with a broad class of models

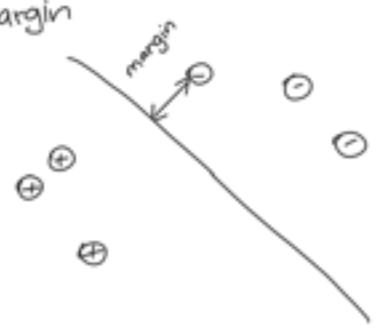
(b) has excellent generalization

<u>Setting</u>: Binary classification: $\hat{y} = \begin{cases} 1 & \text{if } f(x,w) > 0 \\ -1 & \text{else} \end{cases}$

for now, assume separability; also linear case $f(x,w) = w^T x + w_0$

perceptron (hinge loss) → no preference

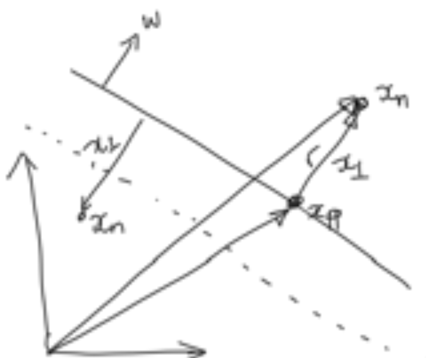
↳ max margin



how far does a pt. need to move before its classification changes?

goal: maximize the minimum margin (later add bases...)

Some geometry:



recall: $w^T x + w_0 = 0$ defines the boundary

$w^T x + w_0 = c$ is parallel to the boundary

$x_n = x_p + x_\perp$

$x_\perp : \underbrace{r}_{\text{margin}} \underbrace{\frac{w}{\|w\|_2}}_{\text{unit vector in the direction of } w}$

$x_p : \underline{w^T x_{\parallel} + w_0 = 0}$

multiply my original eqn by w^T:

$w^T x_n = w^T x_{\parallel} + w^T x_\perp$

$w^T x_n = -w_0 + r \underbrace{\frac{w^T w}{\|w\|_2}}$ $\frac{\|w\|_2^2}{\|w\|_2} = \|w\|_2$

solve for r

$w^T x_n + w_0 = r\,\|w\|_2$

$$\boxed{r = \frac{w^T x_n + w_0}{\|w\|_2}}$$ signed margin

unsigned margin (or just "margin") for some x_n: $y_n\left(\frac{w^T x_n + w_0}{\|w\|_2}\right)$

↳ observe: correctly classified pts will have margin$(x_n) \geq 0$; misclassified pts. will have negative margin.

$$\text{margin}(x_n) = y_n\left(\frac{w^T x_n + w_0}{\|w\|_2}\right)$$

$$\text{margin}(X) = \min_{x_n} \text{margin}(x_n)$$

now, we can write our objective:

$$\max_{w, w_0} \text{margin}(X) = \max_{w, w_0} \min_{x_n} y_n\left(\frac{w^T x_n + w_0}{\|w\|_2}\right)$$

anything misclassified → we'll try to correct it!
even if correctly classified → still try to make margin large!

great idea! looks UGLY!!

let's first notice an over-parametrization in this problem...

$w^T x + w_0 = 0$ as boundary

$$\underbrace{\beta w^T}_{w'^T} x + \underbrace{\beta w_0}_{w_0'} = 0 \qquad \beta > 0$$

$w'^T x + w_0' = 0$

shows up in the margin also:

$$y_n \frac{1}{\beta\|w\|_2} \underbrace{(\beta w^T x + \beta w_0)}_{\text{choose a } \beta \text{ that will achieve this.}} \Leftrightarrow \text{same margin value!}$$

Let's decide on a scaling of w, w_0 s.t. $y_n(w^T x_n + w_0) \geq 1$

$$\max_{w, w_0} \min_{n} \frac{1}{\|w\|_2} y_n(w^T x_n + w_0)$$

$$\max_{w, w_0} \frac{1}{\|w\|_2} \underbrace{\min_{n} y_n(w^T x_n + w_0)}_{1}$$

s.t. $y_n(w^T x_n + w_0) \geq 1$ my constraint

note: optimal sol'n will touch the constraint

$$\max_{w, w_0} \frac{1}{\|w\|_2} \quad \text{s.t. } y_n(w^T x_n + w_0) \geq 1$$

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

$\min \|w\|_2 = \min \|w\|_2^2$
$= \min w^T w$

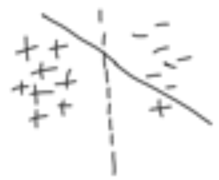
$$\min_{w, w_0} w^T w \quad \text{s.t. } y_n(w^T x_n + w_0) \geq 1$$

very cool because this is a quadratic program with linear constraints.

↳ convex!

↳ very efficient solvers.

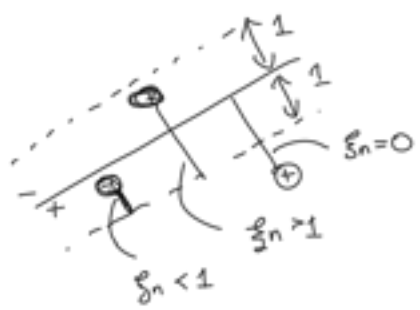
~~bot~~ but what if the data are not separable?



In general, tension btw margin (X) size and correctly classified data/misclassification rate

defined only on correct data

move from "hard margin" to "soft margin"



$\xi_n \geq 0$ measure how much margin violation exists for pt. x_n

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

$$\min_{w, w_0} \frac{1}{2}\|w\|_2^2 + \boxed{C \sum \xi_n}$$
s.t. $y_n(w^T x_n + w_0) \geq 1 - \xi_n$

$\xi_n \geq 0$ -3 .5

"pretend" margin =1, and allow some slack

larger C → enforce constraints more, smaller margin/reg. less

small C → too many errors

usually, we set C w/ cross-validation

question: is this still convex? yes!

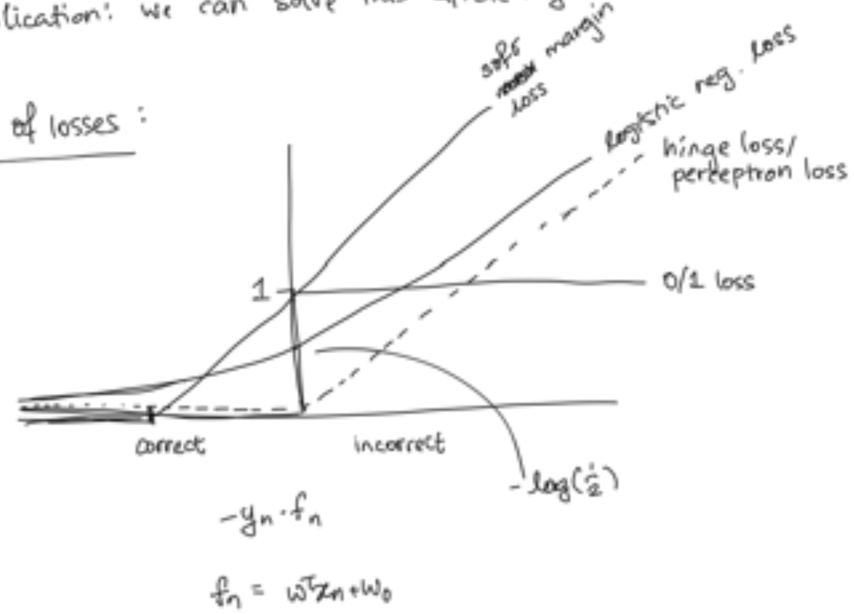
⇒ sum of convex functions is still convex

$$\min_{w, w_0} \frac{1}{2}\|w\|_2^2 + C \sum_n \max(0, \underline{1 - y_n(w^T x_n + w_0)})$$

hinge loss

⇒ implication: we can solve this efficiently!

Review of losses:



the end of our supervised learning unit.

Concept Question: close to the

models: nonparametric $\longrightarrow$ dimencionality & neighbors

parametric: linear, basis, NN

practicals: random forests

losses:

When should you choose each model/loss:

discuss: data? $\xrightarrow[\text{linear vs. nonlin.}]{\text{semi-sup vs. not}}$ size: computation, expressiveness

opt? convex vs. non

access to knowledge? $\longrightarrow$ interpretability

basis vs. nn

type of fit: linear vs. nonlinear

Max margin