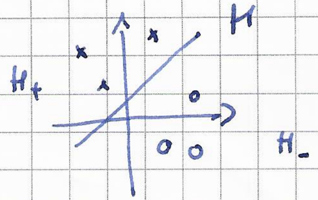
XIII Perceptron

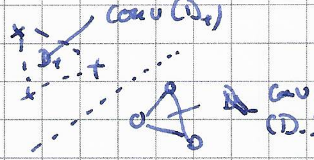
- goal: ~~classify~~ strictly classify binary data $D = \{(x_i, y_i),\ i=1,\dots,n\} \subset \mathbb{R}^d \times \{-1,1\}$
  via hyperplane $H = \{x \in \mathbb{R}^d:\ w \cdot x + b = 0\}$
  and half-spaces $H_\pm = \{x \in \mathbb{R}^d:\ w \cdot x + b \gtrless 0\}$



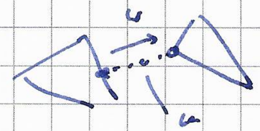
Prop possible only if convex hulls of

$$D_\pm = \{x :\ \cancel{(x,\pm)}\ (x, \pm 1) \in D\}$$

do not intersect.



Pf: "⇒" if $D_+ \subset H_+$ and $D_- \subset H_-$ then $D_+ \cap D_- = \emptyset$
"⇐" and $H_\pm$ convex ~~so~~ so that $\mathrm{conv}(D_\pm) \subset H_\pm$
"⇐": let x_+, x_- be points in $\mathrm{conv}(D_\pm)$ that minimize the distance and set $w = x_+ - x_-$, $m = (x_+ + x_-)/2$
~~and~~ $b = -m \cdot w$



GED

- if compact notation: $\hat{x}_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}$, $\hat{w} = \begin{bmatrix} w \\ b \end{bmatrix}$ so that $H = \{x \in \mathbb{R}^d:\ \hat{x} \cdot \hat{w} = 0\}$
- all points corr. classified iff $y_i\ (\cancel{x_i}\ \hat{w} \cdot \hat{x}_i) > 0$

Alg: given $D \subset \mathbb{R}^d \times \{-1,+1\}$, $\hat{w} \in \mathbb{R}^{d+1}$
~~(1) if~~ while $\exists i \in \{1,\dots,n\}$ s.t. $y_i\ (\hat{w} \cdot \hat{x}_i) \le 0$ ~~set~~ update

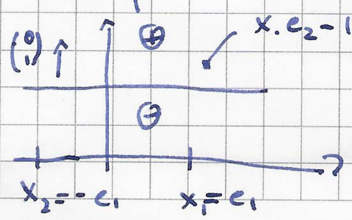
$$\hat{w} = \hat{w} + y_i \hat{x}_i$$

~~otherwise~~ stop
~~(2) repeat (1)~~

- motivation of update

$$y_i\ (\overbrace{\hat{w} + y_i \hat{x}_i}^{\text{new vector}}) \cdot \hat{x}_i = \underbrace{y_i \hat{w} \cdot \hat{x}_i}_{\le 0} + \underbrace{y_i^2 |\hat{x}_i|^2}_{\ge 0}$$

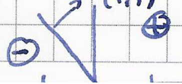
- example



$\hat{w} = \begin{bmatrix} e_2 \\ -1 \end{bmatrix}$

$$y_1 (\hat{w} \cdot \hat{x}_1) = +1 \left(\begin{bmatrix} e_2 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ 1 \end{bmatrix} \right) = -1$$

$$\Rightarrow \hat{w} = \hat{w} + y_1 \begin{bmatrix} e_1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$



<u>Prop</u> (Convergence) if $\hat{w}$ is initialized with $\hat{w}=0$ then alg. terminates within at most R^2/γ^2 steps, provided data is strictly sep. $R = \max |\hat{x}_i|$ $y_i\,\hat{w}^*\cdot\hat{x}_i \geq \gamma > 0$ (proper distance)

<u>Pf</u>: - let correct ~~h~~ sep. hyperplane be given by $\hat{w}^* \in \mathbb{R}^{d+1}$ with ~~...~~ wlog. $|w^*| = 1$

- assume that after $(j-1)$ steps the hyperplane def. by $\hat{w}^{j-1}$ leads to incorrectly ~~labelled~~ classified $\hat{x}$ with label $y=+1$, so that new hyperplane def by $w^j = w^{j-1} + \hat{x}$

- we have $\hat{w}^*\cdot\hat{x} \geq \gamma$ and $\hat{w}^{j-1}\cdot\hat{x} \leq 0$

- now

$$|\hat{w}^j|^2 - |\hat{w}^{j-1}|^2 = |\hat{w}^{j-1}+\hat{x}|^2 - |\hat{w}^{j-1}|^2$$
$$= \underbrace{2\hat{w}^{j-1}\cdot\hat{x}}_{\leq 0} + \underbrace{|\hat{x}|^2}_{\leq R^2} \leq R^2$$
$$\Rightarrow \quad |\hat{w}^j|^2 \leq |\hat{w}^{j-1}|^2 + R^2 \leq |\hat{w}^{j-2}|^2 + 2R^2 \leq \dots \leq \underbrace{|\hat{w}^0|}_{=0} + jR^2$$

- on the other hand $\hat{w}^*\cdot(\hat{w}^j - \hat{w}^{j-1}) = \hat{w}^*\cdot\hat{x} \geq \gamma$

$$|\hat{w}^j| \geq \hat{w}^j\cdot w^* = \hat{w}^{j-1}\cdot\hat{w}^* + \underbrace{\hat{x}\cdot\hat{w}^*}_{\geq\gamma}$$
$$\geq \dots \geq \underbrace{\hat{w}^0\cdot\hat{w}^*}_{=0} + j\gamma$$

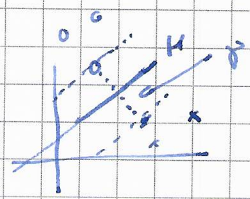
- hence

$$(j\gamma)^2 \leq |\hat{w}^j|^2 \leq jR^2 \quad \Rightarrow \quad j \leq R^2/\gamma^2$$

(qed)

- not difficult to include $w^0 \neq 0$ in proof

Support Vector Machines (SVM)



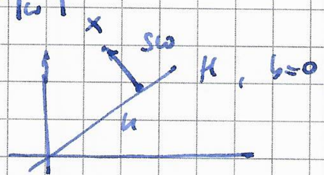
- goal: find sep. hyperplane which max. distances

Def. For $H = \{x \in \mathbb{R}^d : w \cdot x + b = 0\}$ min. distance

$\gamma = \min_{i=1,\dots,n} \operatorname{dist}(x_i, H)$ called **margin**

- hyperplane invariant under scaling, i.e. $H(w,b) = H(\alpha w, \alpha b)$, same pos./neg. half spaces for $\alpha \geq 0$

Lemma $\operatorname{dist}(x, H) = \frac{1}{|w|} | w \cdot x + b | = \left| \frac{w}{|w|} \cdot x + \frac{b}{|w|} \right|$ (*)



Pf: (i) $b = 0$, $|w| = 1$. $x = h + sw$

$h \in H$, $w \cdot h = 0$, $s = \operatorname{dist}(x, H)$
$= |w \cdot x|$

$|w| \neq 1$ same

(ii) $b \neq 0$: translate, scale $\tilde{H} = H - h_0$, $\tilde{x} = x - h_0$

qed

Lemma H separates $D = \{(x_i, y_i) ; i = 1,\dots,n\} \subset \mathbb{R}^d \times \{-1, 1\}$ correctly and strictly iff there exist w, b s.t.

$y_i (w \cdot x_i + b) \geq 1$ f.a. $i = 1,\dots,n$

Pf: "$\Leftarrow$" $y_i (w \cdot x_i + b) \geq 1 \Rightarrow \underbrace{w \cdot x_i + b}_{\neq 0 \text{ an. sign}} \; y_i \quad i = 1,\dots,n$

"$\Rightarrow$" let $\tilde{w}, \tilde{b}$ def. strictly sep. hyperplane, i.e. $\varepsilon > 0$

$\tilde{w} \cdot x_i + \tilde{b} \geq \varepsilon$ if $y_i = 1$, $\tilde{w} \cdot x_i + \tilde{b} \leq -\varepsilon$ if $y_i = -1$

then $y_i (\tilde{w} \cdot x_i + \tilde{b}) \geq \varepsilon$, choose ~~take~~ $w = \tilde{w}/\varepsilon$, $b = \tilde{b}/\varepsilon$

qed

- to find optimal hyperplane max minimal distance using (*)

$$\max_{w,b} \frac{1}{|w|} \quad \text{s.t.} \; \min \; y_i (w \cdot x_i + b) = 1 , \quad i = 1,\dots,n \qquad = M$$

$$\Leftrightarrow \max_{w,b} \frac{1}{|w|} \quad \text{s.t.} \; y_i (w \cdot x_i + b) \geq 1 , \quad i = 1,\dots,n \qquad = M'$$

then $M = M'$: (i) $M' \geq M$ ✓ (ii) $M' = \frac{1}{|w|}$ $\min y_i (w \cdot x_i + b) = r \geq 1$

$= \frac{1}{r} \frac{1}{|w/r|} \leq \frac{1}{|w/r|} \leq M$.

- max pb equ. to min. of reciprocal squared, factor $1/2$

$$\min_{(w,b)\in\mathbb{R}^d\times\mathbb{R}} \frac{1}{2}|w|^2 \quad \text{s.t.} \quad y_i(w\cdot x_i+b)\geq 1,\ i=1,\dots,n$$

<u>Thm</u> If D strictly sep. then exists sol. (w,b) ~~which is unique up to scaling~~ ass. not all points same label.

<u>Pf</u>: (1) adm set. $H=\bigcap H_i$, $H_i=\{\cancel{y_i(w}\ w\cdot x_i+b\geq y_i\}$ is convex; since D sep. there exists adm. pair $(\tilde{w},\tilde{b})$ (after rescaling as above), hence suff to consider (w,b) with $|w|\leq|\tilde{w}|$ and with i_+, i_- s.t. $y_{i_\pm}=\pm 1$

· $b\geq 1-w\cdot x_{i_+}$, $b\leq -1-w\cdot x_{i_-}$,

i.e. consider (w,b) in compact set $\Longrightarrow$ exists minimizer

(2) Since $w\mapsto|w|^2$ is strictly convex, sol $\cancel{b}$ w is unique.
Uniqueness of b: ass. $b_1\neq b_2$, $b_1<b_2$, exists x_1 s.t. $y_1=1$
$y_i(w\cdot x_i+b_1)=1 \Rightarrow y_i(w\cdot x_i+b_2)<1$ ↯ QED

- convex dual pb acc. easier to treat
- incorporate constraint via multipliers $\lambda_i\geq 0$

$$\min_{w,b}\Big(\frac{1}{2}|w|^2+\underbrace{\max_{\lambda_i\geq 0}\sum(-\lambda_i)\big(y_i(w\cdot x_i+b)-1\big)}_{=+\infty \text{ if constraint violated}}\Big)$$

$$=\min_{w,b}\max_{\lambda\geq 0}\ \frac{1}{2}|w|^2-\sum_{i=1}^n\lambda_i y_i\big((w\cdot x_i+b)-1\big)$$

$$\geq\max_{\lambda\geq 0}\min_{w,b}\quad -//-\quad L(w,b;\lambda)$$

for $\lambda\geq 0$ min. (w,b): $\nabla_w L(w,b,\lambda)=0$, $\partial_b L(w,b;\lambda)=0$

$\Rightarrow$ $w-\sum\lambda_i y_i x_i=0$, $\sum\lambda_i y_i=0$

$$=\max_{\lambda\geq 0}\frac{1}{2}\Big|\sum\lambda_i y_i x_i\Big|^2-\Big(\sum\lambda_i y_i x_i\Big)\cdot\underbrace{\Big(\sum\lambda_i y_i x_i\Big)}_{=w}+\sum\lambda_i\cancel{y_i}$$

- can show that equality holds (strong duality)
- points x_i in which dist to this is min. corr. to $w \cdot x + b = \pm 1$ are called supp. vectors, i.e. have $\lambda_i > 0$ (active constraint), opt sep. hyperpl. is SVM
- soft SVMs allow for misclassification by min.:

$$\min_{w, b, \xi} \quad \frac{1}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$