2/20/18

Lecture 6

Null Model

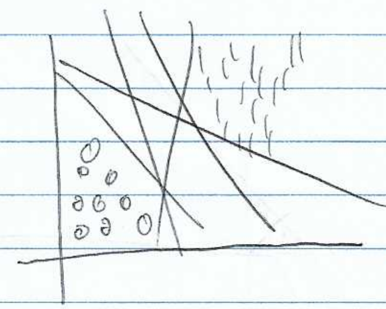
$\vec{x}$'s dont matter

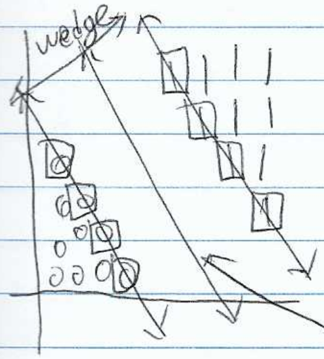
$\mathcal{H} = \{0, 1\}$, $g = \mathcal{A}(\mathbb{D} = \vec{y}, \mathcal{H}) = \text{Mode}[\vec{y}]$

$\mathcal{H} = \{ \mathbb{1}_{\vec{w} \cdot \vec{x} > 0} : \vec{w} \in \mathbb{R}^{p+1} \}$ Assume linear separability

begins with a 1

which hyperplane spec. by $\vec{w}$ is "best"?





(Boxed $\square$ points seem more important than the rest)

"Best" line $\vec{w} \cdot \vec{x} \rightarrow g = \mathbb{1}_{\vec{w} \cdot \vec{x}}$ which is called the "support vector machine"

"best" line $\rightarrow$ "max-margin hyperplane"

"essential observation separation"

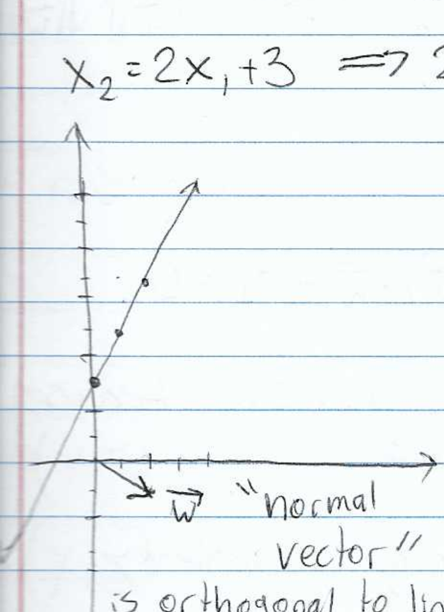
$\mathcal{H} = \{ \mathbb{1}_{\vec{w} \cdot \vec{x} + b > 0} : \vec{w} \in \mathbb{R}^{p}, b \in \mathbb{R} \}$

No longer has a 1 in first column

$x_2 = 2x_1 + 3 \implies 2x_1 - x_2 + 3 = 0 \implies \underbrace{\begin{bmatrix} 2 \\ -1 \end{bmatrix}}_{\vec{w}} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\vec{x}} \underbrace{+3}_{+b} = 0$

$\|\vec{w}\| := \sqrt{\sum_{j=1}^{p} w_j^2}$ "the Euclidean norm"

$\vec{w}_0 = \dfrac{\vec{w}}{\|\vec{w}\|}$ normalized vector $\vec{w}$

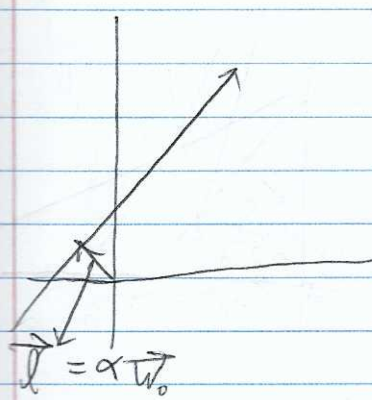


is orthogonal to line

$$\|\vec{w}_0\| = \sqrt{\sum_{j=1}^{p}\left(\frac{w_j}{\|\vec{w}\|}\right)^2} = \sqrt{\frac{1}{\|\vec{w}\|^2}\sum_{j=1}^{p} w_j^2} = \frac{1}{\|\vec{w}\|}\|\vec{w}\| = 1$$

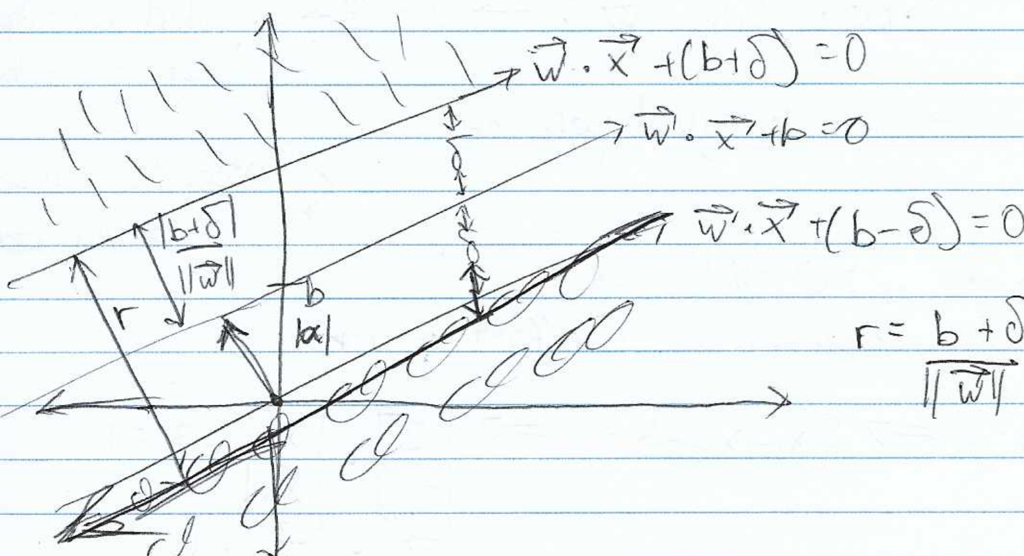
$$\vec{l} = \alpha\vec{w}_0 \Rightarrow \|\vec{l}\| = \sqrt{\sum(\alpha w_{0,j})^2} = \sqrt{\alpha^2 \sum w_{0,j}^2} = |\alpha|\sqrt{\sum w_{0,j}^2} = |\alpha|$$

where $\alpha \in \mathbb{R}$



$\vec{w}\cdot\vec{l} + b = 0$ since $\vec{l}$ is a point on the line

$$\Rightarrow \vec{w}\cdot\vec{l} + b = 0$$
$$\Rightarrow \vec{w}\cdot(|\alpha|\vec{w}_0) + b = 0$$
$$\Rightarrow |\alpha|\,\vec{w}\cdot\frac{\vec{w}}{\|\vec{w}\|} + b = 0$$
$$\Rightarrow |\alpha|\frac{\|\vec{w}\|^2}{\|\vec{w}\|} + b = 0$$
$$\Rightarrow |\alpha| = \left|\frac{-b}{\|\vec{w}\|}\right| = \frac{b}{\|\vec{w}\|}$$



$$r = \frac{b+\delta}{\|\vec{w}\|} - \frac{b-\delta}{\|\vec{w}\|} = \frac{2\delta}{\|\vec{w}\|}$$

$c(\vec{w}\cdot\vec{x} + b) = 0$ s.t. $c \in \mathbb{R}$ $(c\vec{w})\cdot\vec{x} + cb = 0$

coerce $\delta = 1$ (There are no errors in this setup because it is linearly seperable)

minimizing $\|w\|$ means maximizing the margin equivalency

Subject to "no errors"

$\vec{w} \cdot \vec{x} + b + 1 = 0$

$\Rightarrow \vec{w} \cdot \vec{x} + b = -1$

If $\vec{w} \cdot \vec{x}_i + b \geq 1 \Rightarrow y_i = 1 \ \forall i$

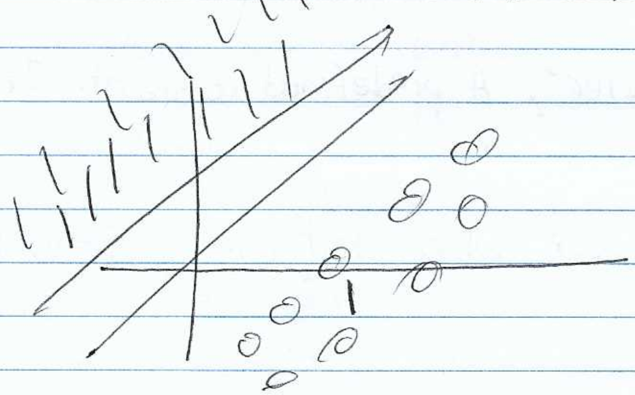
If $\vec{w} \cdot \vec{x}_i + b \leq 1$

$$\boxed{\begin{array}{l} \min \ \|\vec{w}\| \ \text{s.t.} \\ \forall i \ (y_i - \tfrac{1}{2})(\vec{w} \cdot \vec{x}_i + b) \geq -\tfrac{1}{2} \\ \text{over } \vec{w} \in \mathbb{R}^p \quad b \in \mathbb{R} \end{array}}$$

Condition of perfect linear separability

The problem we face is that our data is linearly separable and because it's unrealistic it's not helpful

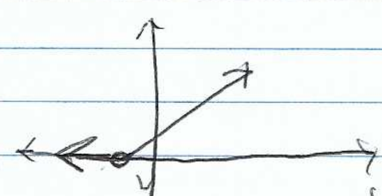
$SAE = \sum_{i=1}^{n} \mathbb{1}_{y_i \neq \hat{y}_i}$ error function, loss function, cost function
Fitness function, objective function



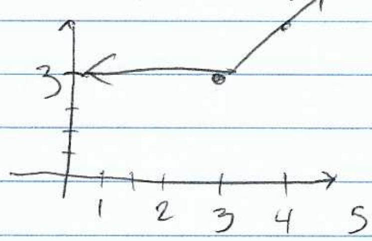
"hinge loss"

$\downarrow$

$H_i := \max\{0, -\tfrac{1}{2} - (y_i - \tfrac{1}{2})(\vec{w} \cdot \vec{x}_i + b)\}$



Ex: $f(x) = \max\{3, x\}$



looks like a "hinge"

If

$(y_i - \tfrac{1}{2})(\vec{w} \cdot \vec{x}_i + b) \geq -\tfrac{1}{2}$

$(y_i - \tfrac{1}{2})(\vec{w} \cdot \vec{x}_i + b) = -\tfrac{1}{2} + d$

$H_i = \max\{-\tfrac{1}{2} - (-\tfrac{1}{2} + d)\}$

$H_i = \max\{0, -d\} = 0$ correct classification

If $(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) < -\frac{1}{2}$ $\quad H_i = \max\{0, -\frac{1}{2} - (-\frac{1}{2} - d)\}$

$(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) = -\frac{1}{2} - d \quad H_i = \max\{0, +d\} = d$

$$\min \underbrace{\frac{1}{n} \sum_{i=1}^{n} H_i}_{\substack{\text{avg hinge loss} \\ \text{"mistake"}}} + \underbrace{\lambda \|\vec{w}\|^2}_{\substack{\text{maximizing the} \\ \text{margin}}} \text{ over } \vec{w} \in \mathbb{R}^p, b \in \mathbb{R}$$

Parameters: $\vec{w}, b$

"λ" is called a "hyperparameter". A predefined constant. It is a timing knob on $\mathcal{A}$.

* perceptron doesn't always work when not linearly separable *