So, lets move on to some hyp. classes.

Linear predictors

Define

$$L_d = \{ h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

$$h_{w,b}(x) = \langle w, x \rangle + b$$

Class of affine functions.

From L_d, we can get different predictors via composition with some $\phi : \mathbb{R} \to \mathcal{Y}$. Further, we can use

$$w' = (b, w_1, \dots w_d) \quad \text{and} \quad x' = (1, x_1, \dots, x_d) \quad \text{to write}$$

$$h_{w,b}(x) = \langle w', x' \rangle$$

$\Rightarrow$ Affine functions in $\mathbb{R}^d$ can be written as homogeneous linear functions in $\mathbb{R}^{d+1}$.

$$f(sx_1, \dots, sx_d) = s^k \cdot f(x_1, \dots, x_d)$$

Here: $k=2$ as $\langle kx, kw \rangle = \sum_{i=1}^{d} kx_i \cdot kw_i$

$$= k^2 \cdot \langle x, w \rangle$$

$$\boxed{\text{if } \phi: \mathbb{R} \to \mathcal{Y} \text{ is } \phi = \text{sign} \text{ we get HALFSPACE classifiers}}$$

$$HS_d = \phi \circ L_d = \{ x \mapsto \text{sign}(h_{w,b}(x)),\ h_{w,b} \in L_d \}$$

VC-dim. of HALFSPACE hypothesis?

$$\mathcal{F} = \{ x \mapsto \operatorname{sign}(\langle w, x \rangle) : w \in \mathbb{R}^d \} \qquad \mathcal{Y} = \{\pm 1\}$$

<u>Theorem</u>: The VC-dim of $\mathcal{F}$ is $VC(\mathcal{F}) = d$!

<u>Proof</u>: we first show $VC(\mathcal{F}) \geq d$ (lower bound)

Let $C = \{c_1, \dots, c_d\}$ where each $c_i \in \mathbb{R}^d$

$$c_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad c_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad c_d = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

at 1st coord   at 2nd   at dth coord

Let

$$w = \begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_d \end{pmatrix}$$ for <u>any</u> given $\{\pm 1\}$ labeling of $\{c_1, \dots c_d\}$

c_{ij} denotes the j-th coord of c_i (and w_j the j-th coord of w)

$$\Rightarrow \langle w, c_i \rangle = \sum_{j=1}^{d} w_j \underbrace{c_{ij}}_{\text{only 1 at } i=j} = w_i = y_i \checkmark$$

We found a set of size <u>d</u> that is shattered by $\mathcal{F}$.

$\Rightarrow$ $VC(\mathcal{F}) \geq d$

We will show the upp. bound $VC(\mathcal{F}) < d+1$ via contradiction!

Assume $C = \{c_1, \dots, c_d, c_{d+1}\}$ is shatt. by $\mathcal{F}$.

That would mean we have $w_1, \dots, w_{2^{d+1}}$ weights that realize the 2^{d+1} possible labelings.

If we write all these labelings in a matrix like

$$\begin{pmatrix} w_1^T c_1 & w_2^T c_1 & \cdots & w_{2^{d+1}}^T c_1 \\ w_1^T c_2 & w_2^T c_2 & & \\ \vdots & \vdots & \ddots & \\ w_1^T c_{d+1} & \vdots & & w_{2^{d+1}}^T c_{d+1} \end{pmatrix}$$

We have $d+1$ rows and 2^{d+1} columns!

We can also write this matrix as a product XW with

$$X = \begin{pmatrix} - & c_1 & - \\ - & c_2 & - \\ & & \\ - & c_{d+1} & - \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} | & | & & | \\ w_1 & w_2 & \cdots & w_{2^{d+1}} \\ | & | & & | \end{pmatrix}$$

$$XW = \begin{pmatrix} \langle w_1, c_1 \rangle & \langle w_2, c_1 \rangle & \cdots \\ \vdots & \ddots & \\ & & \ddots \end{pmatrix}$$

$X \in \mathbb{R}^{(d+1) \times d}$ and $W \in \mathbb{R}^{d \times 2^{d+1}}$

Taking $\operatorname{sign}(XW)$ gives all possible labelings! (see def of ∓)
(in the columns of XW)

Let $M = XW$. We know that

$$\mathrm{rank}(M) \leq \min\left(\mathrm{rank}(X), \mathrm{rank}(W)\right) \leq d$$

as X has only d columns.

<u>Claim:</u> The rows of M are lin. independent (<u>under stoff. ass.</u>)

<u>Proof</u>: Rem. that $v_1, \dots, v_k$ in a vec. space are lin. ind. if

$$a_1 v_1 + a_2 v_2 + \dots a_k v_k = 0$$

is only satisfied <u>iff</u> $\forall i : a_i = 0$!

what is M span?

$$
M = \begin{pmatrix}
w_1^T c_1 & w_2^T c_1 & & w_{2d+1}^T c_1 \\
w_1^T c_2 & w_2^T c_2 & & \\
\vdots & & & \\
w_1^T c_{d+1} & w_2^T c_{d+1} & \cdots & w_{2d+1}^T c_{d+1}
\end{pmatrix}
\begin{array}{l} \to row_1 \\ \\ \\ \to row_{d+1} \end{array}
$$

So,

$$a_1 \cdot row_1 + a_2 \cdot row_2 \;\cdots\; + a_{d+1}\, row_{d+1} =$$

$$a_1 \cdot \left(w_1^T c_1 \quad w_2^T c_1 \quad \cdots \quad w_{2d+1}^T c_1 \right) +$$

$$a_2 \cdot \left(w_1^T c_2 \quad w_2^T c_2 \right.$$

$$\vdots$$

$$a_{d+1} \left(w_1^T c_{d+1} \quad w_2^T c_{d+1} \right.$$

$$\left(\underbrace{a^T \cdot X w_1 \quad a^T X w_2}_{} \quad \text{and so on} \right) \overset{?}{=} 0$$

when

Fact is that there is a k s.t.

$$\text{sign}(a) = \text{sign}(X w_k)$$ due to shattering assumption

So, whatever a is (unless $\vec{0}$), there is always a k s.t. $X w_k$ matches its sign $\Rightarrow a^T X w_k$ = sum of positive numbers and

$$\Big(\cdot \quad \cdot \quad - \quad - \quad \underbrace{-}_{\text{at } \underline{k} \text{ is non-zero}} \quad - \quad - \quad - \quad \cdot \Big) \in \mathbb{R}^{2^{d+1}}$$

$\Rightarrow$ lin. ind. $d+1$ rows ☒

This gives $\boxed{\text{rank}(M) = d+1}$

$\Rightarrow$ CONTRADICTION (implying $VC(\mathcal{F}) < d+1$)

With $VC(\mathcal{F}) \geq d$ and $VC(\mathcal{F}) < d+1$ we have $VC(\mathcal{F}) = d$!

Learning halfspace classifiers

$$HS_d = \{ x \mapsto h_{w,b}(x),\ h_{w,b} \in L_d \}$$

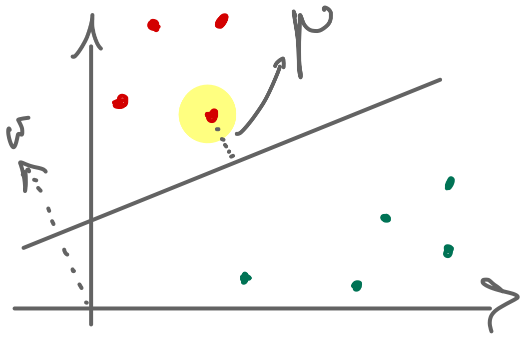
We are going to look at the realizable case!

$$S = ((x_1, y_1), \ldots (x_m, y_m)) \quad \text{Training data} \left(\overset{\text{with}}{Y = \{\pm 1\}} \right)$$

ERM on HS_d means to find $w \in \mathbb{R}^d$ (or $\mathbb{R}^{d+1}$) with bias, s.t.

$$\forall i \in \{1, \ldots, m\} : \ \operatorname{sign}(\langle w, x \rangle) = y_i \quad (*)$$

or, equivalently $\quad y_i \langle w, x_i \rangle > 0$ for all $i \in \{1, \ldots, m\}$



Let $\rho = \min_{i \in \{1,\dots,m\}} y_i \langle w^*, x_i \rangle$

where w^* is a weight vector that satisfies (*).

Letting $\bar{w} = \frac{w^*}{\rho}$, we have

$$\forall i \in \{1,\dots,m\}: \ y_i \langle \bar{w}, x_i \rangle = y_i \langle \frac{w^*}{\rho}, x_i \rangle$$
$$= \frac{1}{\rho} \cdot y_i \langle w^*, x_i \rangle \geq 1$$

This shows that a vector $w \in \mathbb{R}^d$ with

$\forall i: \boxed{y_i \langle w, x_i \rangle \geq 1}$ has to exist! (**)

We can write (**) as $Aw \geq v$ with

$$A = \begin{pmatrix} x_{11} \cdot y_1 & x_{12} \cdot y_1 & \cdots & x_{1d}\, y_1 \\ \vdots & \ddots & & \vdots \\ x_{m1}\, y_m & x_{m2}\, y_m & \cdots & x_{md}\, y_m \end{pmatrix}, \quad w = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix}, \quad v = \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{\text{all ones!}}$$

A linear program (LP) is

$$\underbrace{\max_{w \in \mathbb{R}^d} \langle u, w \rangle}_{\text{linear objective}} \quad \underline{\text{subject to}} \quad \underbrace{Aw \geq v}_{\text{linear inequalities}}$$

In our case, we just set $u \in \mathbb{R}^d$ to some "dummy" vector, e.g.

$$u = (1, \ldots, 1)^T$$