PAC - learning
(Probably Approximately Correct)

Ref: FOMC book by Mohri et al. Chapters 2 & 3

The PAC Learning model:

$\mathcal{X}$: input space (set of all feature vectors)

$\mathcal{Y}$: Set of label e.g., $\{0,1\}$

A concept $c : \mathcal{X} \to \mathcal{Y}$

Concept class $\mathcal{C}$: collection of concepts

Suppose the inputs/examples are picked in an i.i.d. fashion using some distribution D.

The learning problem:

Given $\mathcal{H}$: hypothesis set (not necessarily $= \mathcal{C}$)

& data $S = \{x_1, \dots x_m\} \in$ iid using D,

labels $= \{c(x_1), \dots, c(x_m)\}$,

the goal is to minimize the generalization error,

i.e., $$R(h) = \underset{x \sim D}{P}\left(h(x) \neq c(x)\right) = E\left(\mathbb{1}(h(x) \neq c(x))\right)$$

given hypothesis $h \in \mathcal{H}$

Empirical error:

$$\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(h(x_i) \neq c(x_i))$$

By linearity of expectation,

$$E[\hat{R}_S(h)] = \frac{1}{m}\sum_{i=1}^{m} E(\mathbb{1}(h(x_i) \neq c(x_i)))$$

fixed (not random)

$$= \frac{1}{m}\sum_{i=1}^{m} E_{x\sim D}(\mathbb{1}(h(x) \neq c(x)))$$

$$= E(\mathbb{1}(h(x) \neq c(x)))$$

$$= R(h)$$

PAC-learning:

A concept class $\mathcal{C}$ is PAC-learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that

$\forall\, \epsilon > 0,\ \delta > 0$, for all distributions D on $\mathcal{X}$, and a target concept $c \in \mathcal{C}$, the following holds for any $m \geq poly\left(\frac{1}{\epsilon}, \frac{1}{\delta}, size(c), d\right)$,

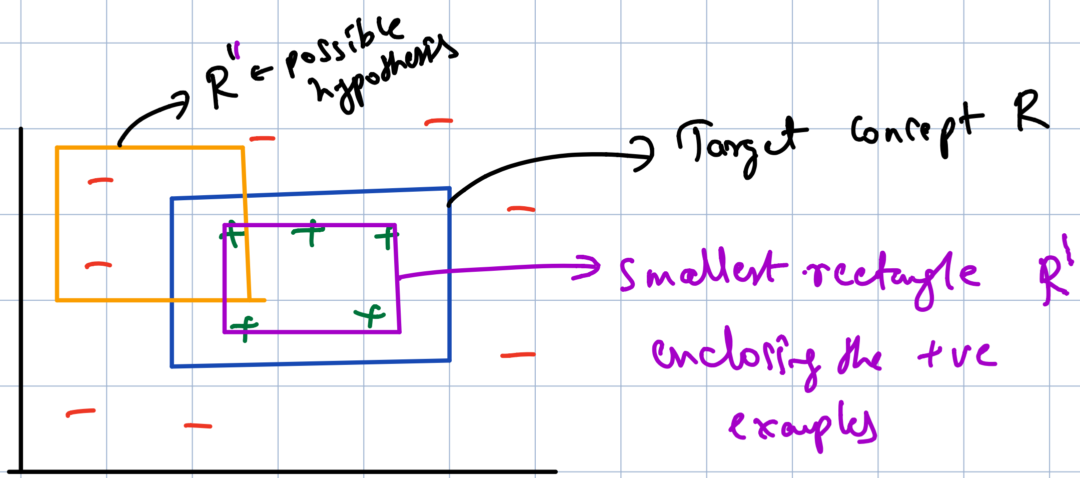
$$P_{S\sim D^m}\left(R(h_S(\mathcal{A})) \leq \epsilon\right) \geq 1-\delta.$$

Note (i) PAC model is distribution-free

(ii) Training & test samples come from the same distribution

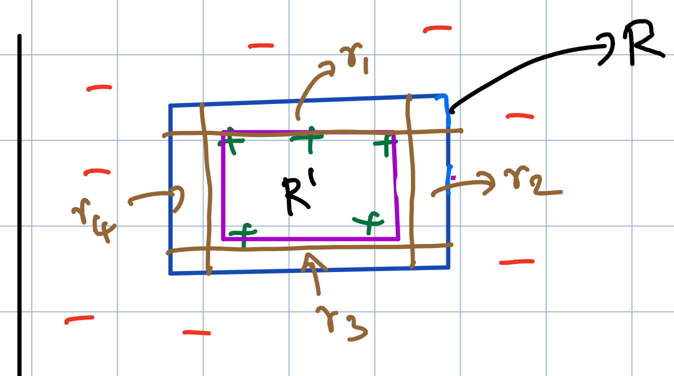
(iii) Learnability → a of the concept class

Example: (Medium-build or axis-aligned rectangles)



Fix $\epsilon > 0$. Let $P(R)$ be the probability mass of R

Suppose that $P(R) > \epsilon$.



$R = [l, r] \times [b, t]$, $r_4 = [l, s_4] \times [b, t]$,

where $s_4 = \min \{ s \mid P([l, s] \times [b, t]) \geq \epsilon/4 \}$

If R' has one side in each r_i, then its error (which is the prob. of region not covered by R') is $< \epsilon$.

Now, If $P(R') > \epsilon$, then R' misses at least one of the regions.

Given data $S = \{x_1, \dots, x_m\}$, let R_S be the smallest rectangle enclosing positive examples.

Then,

$$P(R(R_S) > \epsilon) \leq P\Big[\bigcup_{i=1}^{4} \{R_S \cap r_i = \phi\}\Big)$$

$$\overset{\text{union bound}}{\leq} \sum_{i=1}^{4} P(\{R_S \cap r_i = \phi\})$$

$$\overset{(\text{since } P(r_i) \geq \epsilon/4)}{\leq} 4\left(1 - \frac{\epsilon}{4}\right)^m$$

$$\overset{(\text{used } 1-x \leq e^{-x})}{\leq} 4 \exp\left(-\frac{m\epsilon}{4}\right)$$

To ensure $P(R(R_S) > \epsilon) \leq \delta$,

it is enough if we have

$$4 \exp\left(-\frac{m\epsilon}{4}\right) \leq \delta$$

$$\Updownarrow$$

$$m \geq \frac{4}{\epsilon} \log\left(\frac{4}{\delta}\right)$$

Thus, $\forall \epsilon > 0$, $0 < \delta < 1$, if $m \geq \frac{4}{\epsilon} \log\left(\frac{4}{\delta}\right)$,

then $P(R(R_S) > \epsilon) < \delta$.

or $P(R(R_S) \leq \epsilon) \geq 1-\delta$

Guarantees for finite hypothesis sets - Consistent Case

Theorem: Let $\mathcal{H}$ be a finite set of functions $h: \mathcal{X} \to \mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in \mathcal{H}$, and training data S, returns a "consistent" hypothesis h_S, i.e.,

empirical error $\to \hat{R}_S(h_S) = \frac{1}{m} \sum_{i=1}^{m} 1(h_S(x_i) \neq c(x_i)) = 0$

Then, for any $\epsilon, \delta > 0$,

$$P(R(h_S) \leq \epsilon) \geq 1-\delta \text{ holds if}$$

$$m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right).$$

Proof: Fix $\epsilon > 0$. "Uniform Convergence".

Note: h_S is random as it depends on $S = \{x_1 \ldots x_m\}$

We bound the probability that some $h \in \mathcal{H}$ is consistent and has $R(h) > \epsilon$

Let $\mathcal{H}_\epsilon = \{h \in \mathcal{H} \mid R(h) > \epsilon\}$

Given dataset S,

$$P\left(\hat{R}_S(h)=0\right) \leq (1-\epsilon)^m$$

"fixed h"

$$P\left(\exists h \in \mathcal{H}_\epsilon : \hat{R}_S(h)=0\right)$$

$$= P\left(\hat{R}_S(h_1)=0 \text{ (or) } \hat{R}_S(h_2)=0 \cdots \text{ (or) } \hat{R}_S(h_{|\mathcal{H}_\epsilon|})=0\right)$$

$$\leq \sum_{h\in\mathcal{H}_\epsilon} P\left(\hat{R}_S(h)=0\right)$$

$$\leq \sum_{h\in\mathcal{H}_\epsilon} (1-\epsilon)^m$$

$$\leq |\mathcal{H}|(1-\epsilon)^m$$

$$\leq |\mathcal{H}| \exp(-m\epsilon)$$

equate this to δ & find an expression for m in terms of $|\mathcal{H}|, \epsilon, \delta$.

$$\leq \delta \quad \text{if}$$

$$m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)\right)$$

So, we have

$$P\left(R(h_S) > \epsilon\right) \leq \delta$$

or $$P\left(R(h_S) \leq \epsilon\right) > 1-\delta$$

$\blacksquare$

Equivalent statement:

Set $\delta = |H| \exp(-m\epsilon)$

& find an expression for ϵ, i.e.,

$$\epsilon = \frac{1}{m}\left(\log(|H|) + \log\frac{1}{\delta}\right)$$

$$P\left(R(h_S) \le \frac{1}{m}\left(\log|H| + \log\frac{1}{\delta}\right)\right) > 1-\delta$$

Example: Conjunction of Boolean literals (at most "d" of them)

Boolean literal: x_i or $\bar{x}_i$

Conjunction: $x_1 \wedge \bar{x}_2 \wedge x_4$

0	1	1	0	1	1	+
0	1	1	1	1	1	+
0	0	1	1	0	1	−
0	1	1	1	1	1	+
1	0	0	1	1	0	−
0	1	0	0	1	1	+
0	1	?	?	1	1	$\Rightarrow \bar{x}_1 \wedge x_2 \wedge x_5 \wedge x_6$

Given a dataset, can we find a consistent hypothesis?

Note: negative examples are not informative. So, focus on the positive examples.

Algorithm: Start with all literals, say

$$x_1 \wedge \bar{x}_1 \wedge x_2 \wedge \bar{x}_2 \; - - - \; \wedge x_d \wedge \bar{x}_d$$

& rule out literals that are incompatible with positive examples.

$$|H| = 3^d$$

Using the bound, we get the PAC guarantee for $m \geq \frac{1}{\epsilon}\left(d \log 3 + \log \frac{1}{\delta}\right)$

E.g. for $\delta = 0.02$, $\epsilon = 0.1$, $d = 10$,

the bound is $m \geq 149$, i.e.,

with at least 149 samples, we can obtain a hypothesis that is 90% accurate with probability at least 98%.

Guarantees for a finite hypothesis set, where consistency is not available.

Probability detour: Hoeffding's inequality

Hoeffding's lemma:

A. r.v. $X \in [a,b]$, $EX=\mu$, $b \neq a$

Then, $$E\left(\exp^{\lambda(X-\mu)}\right) \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right)$$

Proof ← skipped.

Hoeffding's inequality:

Let $X_1, \ldots, X_m$ be iid samples, with $X_i \in [a_i, b_i]$. Then, letting $S_m = \sum_{i=1}^{m} X_i$,

$$P\left(S_m - E(S_m) > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)$$

Special case: $X_i \in [a,b]$

$$P\left(S_m - E(S_m) > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{m(b-a)^2}\right)$$

$$\bar{X}_m = \frac{1}{m} S_m, \quad EX_i = \mu$$

$$P\left(\bar{X}_m - \mu > \epsilon\right)$$

$$= P\left(\frac{S_m}{m} - \mu > \epsilon\right) = P\left(S_m - m\mu > m\epsilon\right)$$

$$\leq \exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$$

With Chebyshev's ineq, the RHS is $\frac{(b-a)^2}{m\epsilon^2}$

Proof of Hoeffding's inequality:

to be specified later.

For any r.v. X & $\lambda > 0$, we have

$$P(X \geq \epsilon) = P(\exp(\lambda X) \geq \exp(\lambda \epsilon))$$

Markov ineq.

$$\leq \frac{E(\exp(\lambda X))}{\exp(\lambda \epsilon)}$$

$$P(S_m - ES_m \geq \epsilon)$$

$$\leq e^{-\lambda \epsilon} E(\exp(\lambda(S_m - ES_m)))$$

Samples are independent

$$\leq e^{-\lambda \epsilon} \prod_{i=1}^{m} E(\exp(\lambda(X_i - EX_i)))$$

Hoeffding's lemma

$$\leq e^{-\lambda \epsilon} \prod_{i=1}^{n} \exp\left(\frac{\lambda^2 (b_i - a_i)^2}{8}\right)$$

$$= e^{-\lambda \epsilon} \exp\left(\lambda^2 \sum_{i=1}^{m} \frac{(b_i - a_i)^2}{8}\right)$$

Set $\lambda = \frac{4\epsilon}{\sum_{i=1}^{m}(b_i - a_i)^2}$

$$\leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)$$

$\blacksquare$

Using a parallel argument, one can obtain

$$P\left(S_m - ES_m \leq -\epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)$$

Combining, we have

$$P\left(|S_m - ES_m| > \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)$$

<u>Back to PAC-learning with finite hypothesis sets —</u>

<u>the "inconsistent" case</u>

Using Hoeffding's inequality, for any "fixed" hypothesis $h: \mathcal{X} \to \{0,1\}$, we have

over the distribution of S that has "m" iid samples

$$P\left(\hat{R}_S(h) - R(h) \geq \epsilon\right) \leq \exp\left(-2m\epsilon^2\right)$$

$$P\left(\hat{R}_S(h) - R(h) \leq -\epsilon\right) \leq \exp\left(-2m\epsilon^2\right)$$

(or)

$$P\left(|\hat{R}_S(h) - R(h)| \geq \epsilon\right) \leq 2\exp\left(-2m\epsilon^2\right)$$

<u>Example:</u> (Tossing a coin)

A coin with bias p (=prob(heads)) is tossed "m" times. iid hypothesis $h =$ predict tails.

$$R(h) = p \quad , \quad \hat{R}_S(h) = \hat{p} \rightarrow \text{sample avg of heads}$$

$$P(|\hat{p} - p| \geq \epsilon) \leq 2\exp(-2m\epsilon^2)$$

equate to δ

$$P\left(|\hat{p} - p| \geq \sqrt{\frac{\log 2/\delta}{2m}}\right) \leq \delta$$

$$P\left(|\hat{p} - p| \leq \sqrt{\frac{\log 2/\delta}{2m}}\right) \geq 1-\delta$$

E.g. $\delta = 0.02$, $m = 500$. Then,

With probability at least 98%,

$$|\hat{p} - p| \leq \sqrt{\frac{\log(100)}{1000}} \approx 0.067$$

Claim: Let $\mathcal{H}$ be a finite hypothesis set.

Then, for any $\delta > 0$, w.p. $(1-\delta)$, we have

$$\forall h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log|\mathcal{H}| + \log 2/\delta}{2m}}$$

S is a dataset with "m" iid samples

Remark: In comparison to the bound in the consistent case, we have a bound on that scales with $\sqrt{\frac{\log|\mathcal{H}|}{m}}$ in the inconsistent case.

In the consistent case, the bound scaled as "$\frac{\log|\mathcal{H}|}{m}$".

<u>Proof:</u> Let $h_1, \ldots, h_{|\mathcal{H}|}$ be the elements of $\mathcal{H}$.

Then,

$$P\left(\exists h \in \mathcal{H} \text{ s.t. } |\hat{R}_S(h) - R(h)| > \epsilon\right)$$

$$= P\big(|\hat{R}_S(h_1) - R(h_1)| > \epsilon \text{ (or)}$$

$$|\hat{R}_S(h_2) - R(h_2)| > \epsilon \text{ (or)}$$

$$\vdots$$

$$|\hat{R}_S(h_{|\mathcal{H}|}) - R(h_{|\mathcal{H}|})| > \epsilon \big)$$

$$\leq \sum_{h \in \mathcal{H}} P\left(|\hat{R}_S(h) - R(h)| > \epsilon\right)$$

$$\leq 2|\mathcal{H}| \exp(-2m\epsilon^2)$$

To complete the proof, set

$$\delta = 2|\mathcal{H}| \exp(-2m\epsilon^2)$$

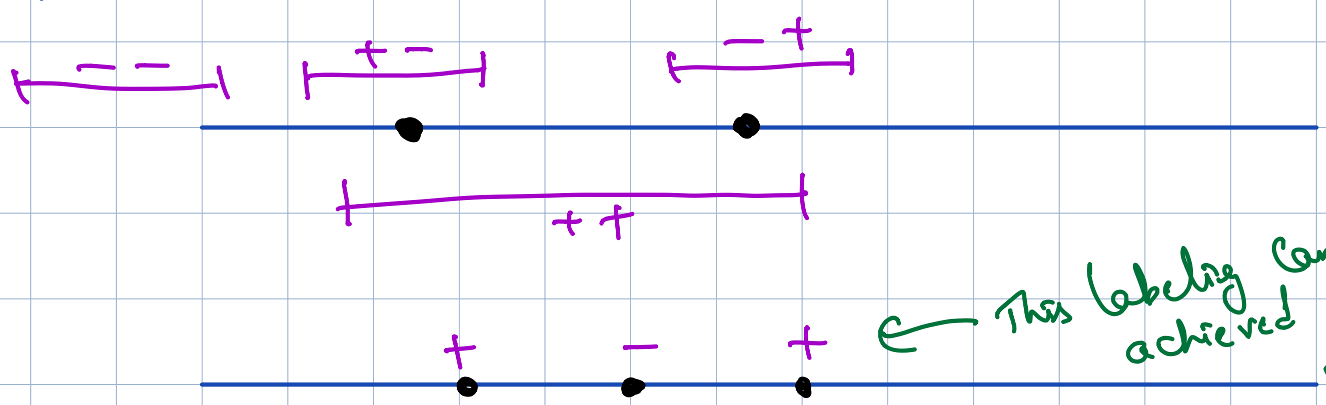
$\blacksquare$

The Case of infinite hypothesis sets:
Growth function & VC-dimension

<u>Growth function</u> $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis set $\mathcal{H}$ is defined as

$$\forall m \in \mathbb{N}, \quad \Pi_{\mathcal{H}}(m) = \max_{\{x_1, \dots, x_m\} \subseteq X} \left| \{ h(x_1) \dots h(x_m) \mid h \in \mathcal{H} \} \right|$$

<u>Example:</u> Intervals on a real line



← This labeling cannot be achieved by an interval

<u>Claim:</u> " Growth function generalization bound "

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, w.p. $(1-\delta)$

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log 1/\delta}{2m}}$$

VC - dimension:

Shattering: A set S of m points is said to be shattered by a hypothesis set $\mathcal{H}$ if

$$\Pi_{\mathcal{H}}(m) = 2^m.$$

Def: The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be shattered by $\mathcal{H}$.

$$VCdim(\mathcal{H}) = \max\{m \mid \Pi_{\mathcal{H}}(m) = 2^m\}$$

Remark:

① If $VC\dim(\mathcal{H}) = d$, then there exists a set of size d that can be shattered by $\mathcal{H}$.

This does not imply that all sets of size "d or less than d" are shattered by $\mathcal{H}$.

(2) To give a lower bound for $VCdim(\mathcal{H})$, it is enough to show a set of size d that can be shattered by $\mathcal{H}$ i.e.,

$$VC\dim(\mathcal{H}) \geq d$$

For the upper bound, one has to show that no set S of cardinality "$d+1$" can be shattered by $\mathcal{H}$, i.e.,

$$VC\dim(\mathcal{H}) < d+1$$

<u>Example 1:</u> "Intervals on the real line"

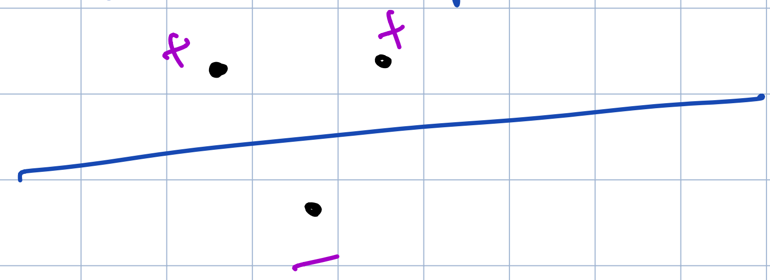
$$VC\dim(\text{"set of intervals"}) = 2.$$

<u>Example 2:</u> "Hyperplanes".

Consider the set of hyperplanes in $\mathbb{R}^2$

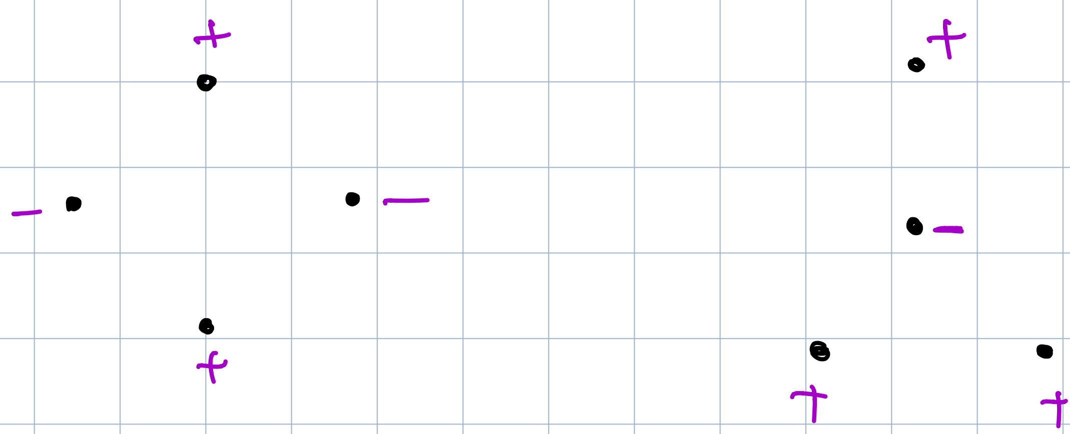
$$VC\dim(\text{hyperplanes in } \mathbb{R}^2) \geq 3$$

Set S = "3 non-collinear points"



& so on

To show $VC(\dim(\text{hyperplanes}) < 4$



So, $VC \dim(\text{hyperplanes}) = 3.$

This result can be generalized to $\mathbb{R}^d$, i.e.,

$$VC \dim(\text{hyperplanes in } \mathbb{R}^d) = d+1$$

Lower bound: Need a set S of "d+1" points that can be shattered.

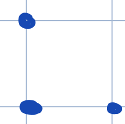
$$S = \{x_0, x_1, \ldots x_d\} \qquad x_i \in \mathbb{R}^d$$

$x_0 =$ Origin $= (0, \dots, 0)$

$x_1 = e_1 = (1, 0, \dots, 0)$

$\vdots$

$x_d = e_d = (0, \dots, 0, 1)$



$y_0, \dots, y_d \in \{+1, -1\}$ be the labeling

$w = (y_1, \dots, y_d)$

hyperplane for classification:

$$w^T x + \frac{y_0}{2} \quad \text{, where} \quad x \in \mathbb{R}^d$$

Want: $\quad \operatorname{sgn}\left(w^T x_i + \frac{y_0}{2}\right) = y_i$

The above holds because

$$\operatorname{sgn}\left(w^T x_i + \frac{y_0}{2}\right)$$

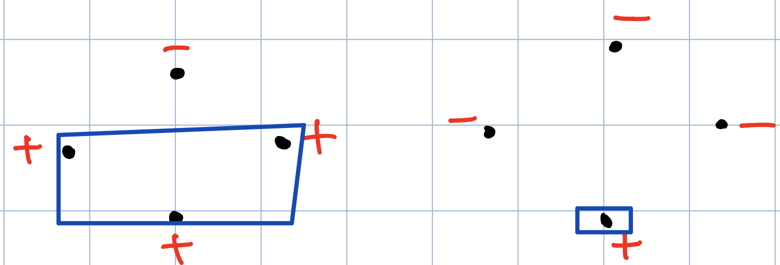
$$= \operatorname{sgn}\left(y_i + \frac{y_0}{2}\right) = y_i \qquad \text{for } i = 1 \dots d$$

One can show an upper bound of $d+2$ for VCdim(hyperplanes in $\mathbb{R}^d$) using "Radon's theorem" (check the textbook).

$$\text{VCdim}(\text{hyperplanes in } \mathbb{R}^d) = d+1$$

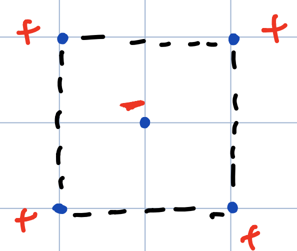
<u>Example 3:</u> Axis-aligned Rectangles. (AAR)

$$\text{VCdim}(\text{AAR}) \geq 4$$



& so on

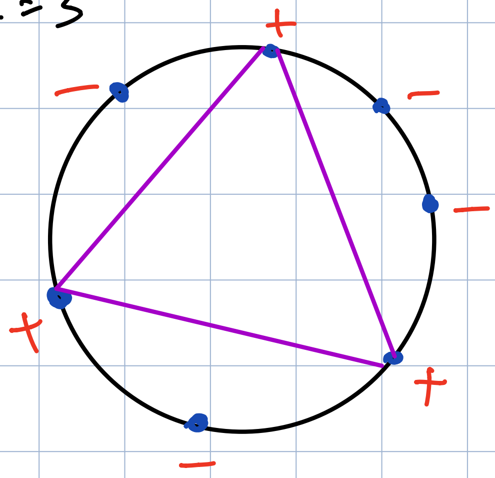
$$\text{VCdim}(\text{AAR}) < 5$$



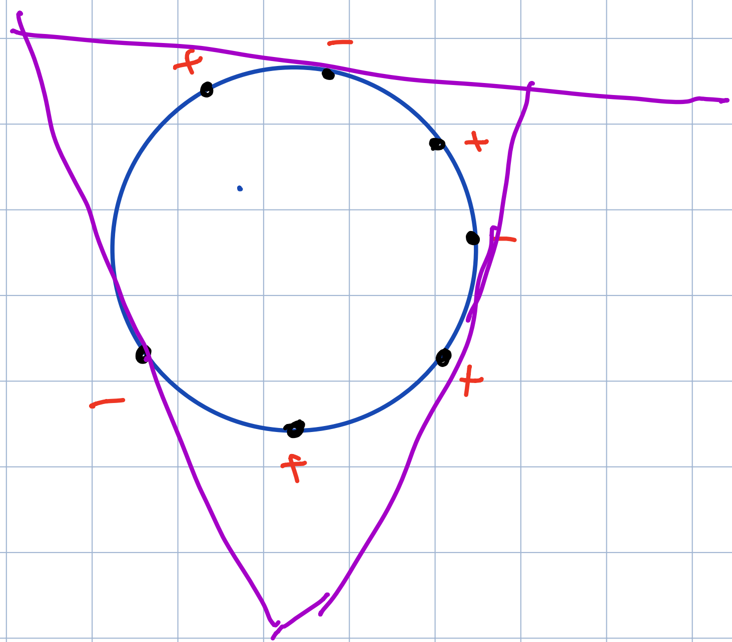
<u>Example:</u> "Convex d-gons in the plane"

Claim: $\text{VCdim}(\text{Convex } d\text{-gons}) \geq 2d+1$

E.g. $d = 3$



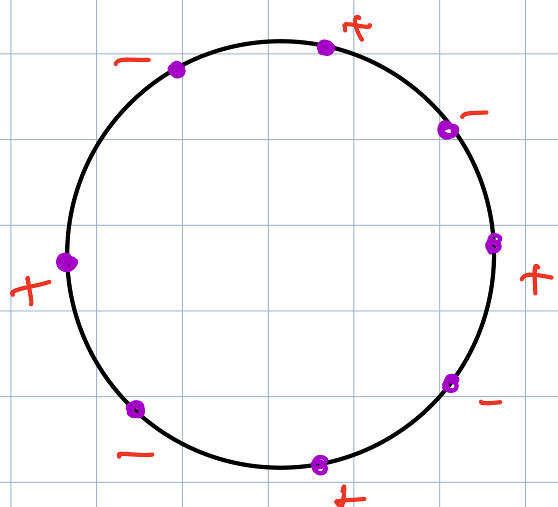
$\leftarrow$ # +ves $<$ # -ves



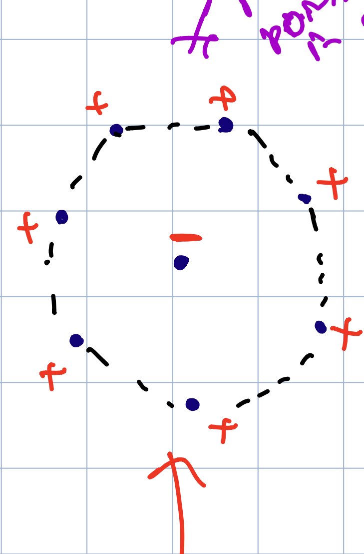
$\leftarrow$ # +ves $>$ # -ves

Upper bound

$$\text{VCdim}(d\text{-gons}) \le 2d+2$$



If this pattern of points occurs in data



Case where no point is in the convex hull of remaining points.

a point is in the interior of the rest of the points (= convex hull of the remaining points)

Hence, $VCdim(d\text{-gons in the plane}) = 2d+1$

<u>Note</u>: $VCdim(\text{Convex polygons}) = +\infty$.

(or) $\Pi_{\text{convex polygons}}(m) = 2^m \quad \forall m \geq 1$

(Sauer's lemma): Let $\mathcal{H}$ be a hypothesis set with $VCdim(\mathcal{H}) = d$. Then, $\forall m \geq 1$,

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

"No proof".

<u>Corollary</u>: Let $\mathcal{H}$ be a hypothesis set with $VCdim(\mathcal{H}) = d$.

Then, $\forall m \geq d$,

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$$

Pf: From Sauer's lemma,

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

$$\leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i}$$

$$\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i}$$

$$= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^{i}$$

$$= \left(\frac{m}{d}\right)^{d} \left(1+\frac{d}{m}\right)^{m}$$

$$\leq \left(\frac{m}{d}\right)^{d} e^{d}$$

Note:- Two Cases

$\rightarrow$ $VC\dim(\mathcal{H}) = d < \infty \Rightarrow \Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^{d}$

$\rightarrow$ $VC\dim(\mathcal{H}) = \infty \Rightarrow \Pi_{\mathcal{H}}(m) = 2^{m}$

Main claim $\Rightarrow$ Generalization bound with VC-dimension

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ & $VC\dim(\mathcal{H}) = d$.

For any $\delta > 0$, w.p. $(1-\delta)$, $\forall h \in \mathcal{H}$,

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log 1/\delta}{2m}}$$