MACHINE LEARNING – VO 2020

KWITT ROLAND

Uniform Convergence

$S \sim D^m$ ↓ over $X \times Y$ —(1)→ LEARNER —(2)→ Selects $h \in H$ by ERM (h_S)

Good property: Ideally $L_S(h_S)$ should be close to $L_D(h_S)$ – this should hold for any $h \in H$.

Def. A sample S is called **ε-representative** (with respect to D, H, loss function ℓ and $Z = X \times Y$) if

$$\forall h \in H : |L_S(h) - L_D(h)| \leq \varepsilon$$

Lemma: If a training sample, S, is **$\frac{\varepsilon}{2}$-representative** then any output of $ERM_H(S)$, say h_S, satisfies

$$L_D(h_S) \leq \min_{h \in H} L_D(h) + \varepsilon$$

Proof:

$$L_D(h_S) \leq L_S(h_S) + \varepsilon/2 \quad \text{(by } \varepsilon/2\text{-rep.)}$$

$$\leq L_S(h) + \varepsilon/2 \quad \text{(as } h_S \text{ is a ERM hyp.)}$$

$$\leq L_D(h) + \varepsilon/2 + \varepsilon/2 \quad \text{(by } \varepsilon/2\text{-rep.)}$$

$$= L_D(h) + \varepsilon$$

Since this inequality chain holds for any $h \in H$, it also holds for the minimum.

$$\Rightarrow L_D(h_S) \leq \min_{h \in H} L_D(h) + \varepsilon$$

□

Def. A hyp. class H has the uniform convergence property, if $\exists\, m_H^{UC} : (0,1)^2 \to \mathbb{N}$, such that for every $\varepsilon, \delta \in (0,1)$ and every prob. distribution D over $Z = X \times Y$, if S is a sample of size $m \geq m_H^{UC}(\varepsilon,\delta)$, then, with prob. of at least $1-\delta$, S is ε-representative.

We already know that finite hyp. classes (H) are PAC learnable via ERM.

Question: Are finite hyp. classes agnostic PAC learnable?

Let $S = (z_1, \dots z_m)$, $z_i \sim \mathcal{D}$, $z_i \in X \times Y$ (the z_i are drawn i.i.d from $\mathcal{D}$.

$$\mathcal{D}^m\left(\{S : \forall h \in H, |L_S(h) - L_{\mathcal{D}}(h)| \le \varepsilon\}\right) \ge 1-\delta$$

Equivalently,

$$\mathcal{D}^m\left(\{S : \exists h \in H, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right) < \delta$$

$$\mathcal{D}^m\left(\bigcup_{h \in H} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right) \le$$

Union bound

$$(*) \quad \sum_{h \in H} \mathcal{D}^m\left(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right)$$

$L_S(h) = \frac{1}{m} \cdot \sum_{i=1}^{m} \ell(h, z_i)$

Taking the expectation, gives us $L_{\mathcal{D}}(h)$

$$L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}}\left[\ell(h, z)\right]$$

If $\ell = \ell_{0-1}$ (0-1 loss), we know the lower and upper bound $(0, 1)$.

By the Hoeffding inequality, we get

$$\mathcal{D}^m\left(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right) \le 2 \cdot e^{-\varepsilon^2 m / (b-a)^2} \qquad \begin{vmatrix} b=1 \\ a=0 \\ (b-a)=1 \end{vmatrix}$$

$$= 2 \cdot e^{-\varepsilon^2 m}$$

Overall $(*)$, we get

$$\mathcal{D}^m\left(\{S : \exists h \in H, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right) \le \sum_{h \in H} 2 \cdot e^{-\varepsilon^2 m} = 2 \cdot |H| \cdot e^{-\varepsilon^2 m}$$

For $2 \cdot |H| \cdot e^{-\varepsilon^2 m}$ to be smaller than $\delta \in (0,1)$, we need

$$m > \log\left(\frac{2 \cdot |H|}{\delta}\right) \cdot \frac{1}{2\varepsilon^2}$$

Corollary: Let H be a finite hyp. class ($|H| < \infty$) and $\ell : H \times Z \to [0,1]$. Then H has the uniform convergence property with

$$m_H^{UC}(\varepsilon, \delta) \leq \left\lceil \log\left(\frac{2 \cdot |H|}{\delta}\right) \cdot \frac{1}{2\varepsilon^2} \right\rceil$$

Furthermore, H is agnostic PAC learnable by ERM with

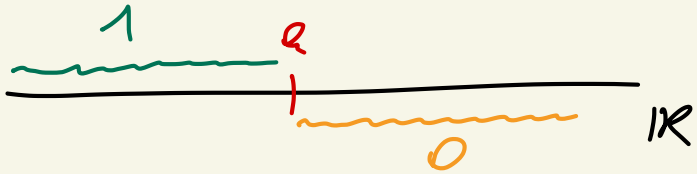
$$m_H(\varepsilon, \delta) \leq m_H^{UC}\left(\frac{\varepsilon}{2}, \delta\right) \leq \left\lceil 2 \cdot \log\left(\frac{2 \cdot |H|}{\delta}\right) / \varepsilon^2 \right\rceil$$

Last lecture: Finite H are agnostic PAC learnable.
(shown via uniform convergence).

Example (of an infinite hyp. class that is PAC learnable).

(*) $H = \{h_a : a \in \mathbb{R}\}$, $h_a : \mathbb{R} \to \{0,1\}$

$$h_a(x) = \mathbb{1}_{[x<a]} = \begin{cases} 1, & \text{if } x < a \\ 0, & \text{else} \end{cases}$$



("Class of threshold functions")

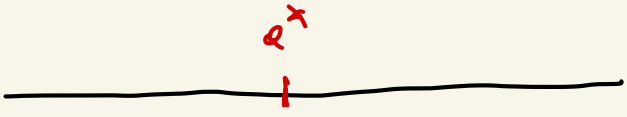
Lemma: Let H be the hyp. class (*) of threshold functions.
Then, H is PAC learnable (via ERM) with sample complexity

$$m_H(\varepsilon, \delta) \leq \left\lceil \frac{\log\left(\frac{2}{\delta}\right)}{\varepsilon} \right\rceil .$$

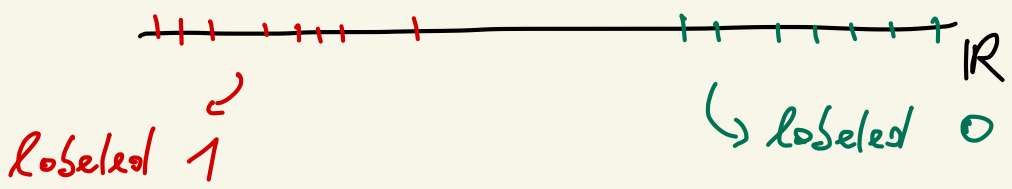
Proof: Let h^* be such that

$$L_{D,f}(h^*) = 0$$

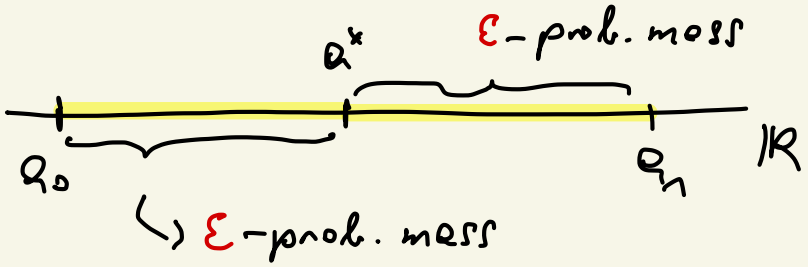
and let $a^* \in \mathbb{R}$ be the corresponding threshold a^*.



Let S be our training data.



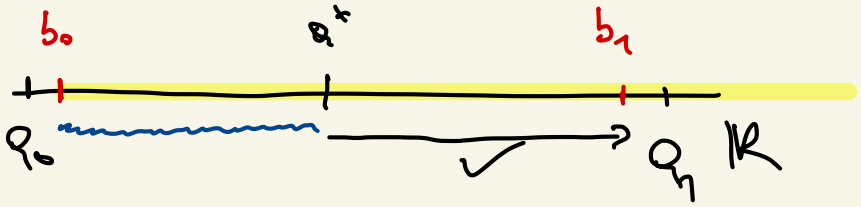
$$b_0 = \max\{x : (x,1) \in S\}$$
$$b_1 = \min\{x : (x,0) \in S\}$$



We have $a_0 < a^* < a_1$ such that

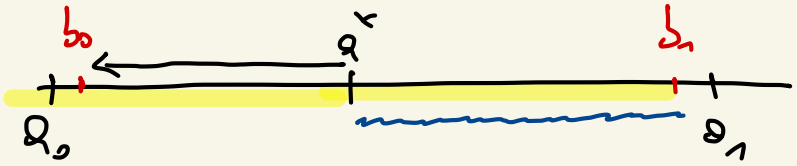
$$\mathcal{D}(\{x : x \in (a_0, a^*)\}) = \mathcal{D}(\{x : x \in (a^*, a_1)\}) = \varepsilon$$

Case 1)



Say b_0 is the thr. of the ERM hypothesis.

Case 2)



ERM picks b_1 as the thr.

$\Rightarrow$ It is sufficient that $b_0 \geq a_0$ AND $b_1 \leq a_1$ for

$$L_{D,f}(h_S) \leq \varepsilon$$

with h_S being an ERM hypothesis. We know that the thr. b_S of h_S has to be within (b_0, b_1), ie., $b_S \in (b_0, b_1)$.

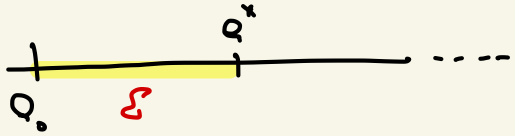
$$\Rightarrow \underset{S|x \sim D^m}{\mathbb{P}}\left[L_{D,f}(h_S) > \varepsilon\right] \leq \underset{S|x \sim D^m}{\mathbb{P}}\left[b_0 < a_0 \text{ OR } b_1 > a_1\right]$$

$$b_0 = \max\{x : (x,1) \in S\}$$

$$\overset{\text{U.B.}}{\leq} \mathbb{P}\left[b_0 < a_0\right] + \mathbb{P}\left[b_1 > a_1\right]$$

By construction (of a_0, a_1), we have

$$\mathbb{P}\left[b_0 < a_0\right] = D^m\left(\{x : (x,y) \in S, x \notin (a_0, a^*)\}\right) = (1-\varepsilon)^m$$



$$\Rightarrow \mathbb{P}\left[b_0 < a_0\right] = (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

Combined, we get: $\mathbb{P}\left[L_{D,f}(h_S) > \varepsilon\right] \leq \underbrace{2e^{-\varepsilon m}}_{< \delta} \Rightarrow m > \frac{\log\left(\frac{2}{\delta}\right)}{\varepsilon}$

23

No Free-Lunch Theorem

Learning task is defined via a distribution over $X \times Y$ (in our case $X \times \{0,1\}$). Goal: Find a predictor

$$h: X \to Y = \{0,1\}$$

that has small risk, i.e.,

$L_D(h)$ is small!

Question: Is prior knowledge really necessary?

So far, prior knowledge came in the form of

1. realizability ($\exists h$ in some H, with $L_D(h) = 0$)
2. assuming $\min_{h \in H} L_D(h)$ is small

(3. assuming something about D)

Thm: Let A be any learning algorithm for the task of binary classification (under 0-1 loss) over a domain X. Let m be any number smaller than $|X|/2$, representing the training set size, i.e., $|S| = m$. Then, there exists a distribution D over $X \times Y = \{0,1\}$, such that

1. $\exists f: X \to \{0,1\}$ with $L_D(f) = 0$!
2. with prob. of at least $\frac{1}{7}$ over the choice of $S \sim D^m$, we have $L_D(A(S)) \geq \frac{1}{8}$!

wrt. 1, we could have, e.g. $H = \{f\}$, or H that is finite and contains f.

Proof: Let X be our domain and $C \subset X$, such that $|C| = 2m$.

We see that there are 2^{2m} functions from $C \to \{0,1\}$.

Let $f_1, \dots, f_T$ denote these functions, $T = 2^{2m}$.

Define:

$$D_i(\{(x,y)\}) = \begin{cases} \frac{1}{|C|}, & \text{if } y = f_i(x) \\ 0, & \text{else} \end{cases}$$

$\Rightarrow L_{D_i}(f_i) = 0$ by construction!

We show the following: For every algorithm A, receiving training data of size m from $X \times \{0,1\}$, it holds that

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m} \left[L_{D_i}(A(S)) \right] \geq \frac{1}{4}$$

Consequence: For every algorithm A', receiving m training samples from $X \times \{0,1\}$, there exists a function $f: X \to \{0,1\}$ and a distribution D over $X \times \{0,1\}$ such that $L_D(f) = 0$ and

$$\mathop{\mathbb{E}}_{S \sim D^m} \left[L_D(A'(S)) \right] \geq \frac{1}{4}$$

Notation: $[T] = \{1, \dots, T\}$

Our training data, S, is $S = (x_1, \dots, x_m)$. Since $|C| = 2m$, there are $(2m)^m$ possible training sequences from C.

Denote these sequences by $S_1, \dots, S_k$ with $k = (2m)^m$.

If S_j is labeled by f_i, we write S_j^i.

$$S_j^i = \big((x_1, f_i(x_1)), \dots, (x_m, f_i(x_m))\big)$$

We can write

$$\mathop{\mathbb{E}}_{S \sim D_i^m}\big[L_{D_i}(A(S))\big] = \frac{1}{k} \cdot \sum_{j=1}^{k} L_{D_i}\big(A(S_j^i)\big)$$

$$\Rightarrow \max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m}\big[L_{D_i}(A(S))\big] = \max_{i \in [T]} \frac{1}{k} \cdot \sum_{j=1}^{k} L_{D_i}\big(A(S_j^i)\big) \quad (*)$$

$$(*) \quad \max_{i \in [T]} \frac{1}{k} \cdot \sum_{j=1}^{k} L_{D_i}\big(A(S_j^i)\big) \geq \frac{1}{T} \cdot \sum_{i=1}^{T} \frac{1}{k} \cdot \sum_{j=1}^{k} L_{D_i}\big(A(S_j^i)\big)$$

$$= \underbrace{\frac{1}{k} \cdot \sum_{j=1}^{k} \frac{1}{T} \cdot \sum_{i=1}^{T} L_{D_i}\big(A(S_j^i)\big)}_{\text{avg. over } j = 1, \dots, k}$$

$$\geq \min_{j \in [k]} \frac{1}{T} \cdot \sum_{i=1}^{T} L_{D_i}\big(A(S_j^i)\big)$$

output is a hyp. $h: X \to \{0,1\}$

Next step: Fix $j \in [k]$.

if $S = (x_1, \dots, x_m)$ and $|C| = 2m$, there are $\rho \geq m$ remaining instances in C, denoted by $v_1, \dots, v_\rho$.

For any $h : C \to \{0,1\}$, we have that

$$L_{D_i}(h) = \frac{1}{2m} \cdot \sum_{x \in C} \mathbb{1}_{h(x) \neq f_i(x)}$$

$$\geq \frac{1}{2m} \sum_{r=1}^{\rho} \mathbb{1}_{h(v_r) \neq f_i(v_r)}$$

$$\geq \frac{1}{2\rho} \sum_{r=1}^{\rho} \mathbb{1}_{h(v_r) \neq f_i(v_r)} \quad (**)$$

By using $(**)$, we get that

$$\frac{1}{T} \cdot \sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right) \geq \frac{1}{T} \cdot \sum_{i=1}^{T} \cdot \frac{1}{2\rho} \cdot \sum_{r=1}^{\rho} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)}$$

$$= \frac{1}{2\rho} \cdot \sum_{r=1}^{\rho} \cdot \frac{1}{T} \cdot \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)}$$

$$= \frac{1}{2} \cdot \frac{1}{\rho} \cdot \sum_{r=1}^{\rho} \frac{1}{T} \cdot \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)}$$

(avg. over $r = 1, \dots \rho$)

$$(***) \quad \geq \frac{1}{2} \cdot \min_{r \in [\rho]} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)}$$

Next step: Fix $r \in [p]$ $\quad [p] = \{1, \dots, p\}$

C is of size $2m$.

Say $C = 4 = 2m \Rightarrow m = 2$

		$f_1, f_2, \dots$	f_T
	x_1	0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1	
	x_2	0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1	
$v_r =$	x_3	0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1	
	x_4	0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1	

In general, we can partition all functions into $T/2$ disjoint pairs. For each pair $(f_i, f_{i'})$ we have that for every $c \in C$, $f_i(c) \neq f_{i'}(c) \Leftrightarrow c = v_r$.

It follows that $S_j^i = S_j^{i'}$ for each pair!

$$\Rightarrow \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)} + \mathbb{1}_{A(S_j^{i'})(v_r) \neq f_{i'}(v_r)} = 1$$

one of them has to be 1, the other 0!

We get that

$$\frac{1}{T} \cdot \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)} = \frac{1}{2}$$

With (***), we have

$$\frac{1}{T} \cdot \sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right) \geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

In combination:

$$\max_{i \in [T]} \underbrace{\frac{1}{k} \cdot \sum_{j=1}^{k} L_{D_i}\left(A(S_j^i)\right)}_{\mathbb{E}_{S \sim D_i^m}\left[L_{D_i}(A(S))\right]} \geq \frac{1}{4}$$

$\Rightarrow$ This shows that

$$\max_{i \in [T]} \mathbb{E}_{S \sim D_i^m}\left[L_{D_i}(A(S))\right] \geq \frac{1}{4}$$

and we know that this implies (first page) for any A' receiving m samples from $X \times \{0,1\}$ there exists a function $f: X \to [0,1]$ and distribution D over $X \times \{0,1\}$ such that $L_D(f) = 0$ and

$$\mathbb{E}_{S \sim D^m}\left[L_D(A'(S))\right] \geq \frac{1}{4}$$

(xxxx)

Lemma (related to the MARKOV inequality). Let Z be a random variable that takes on values in $[0,1]$. Assume $\mathbb{E}[Z] = \mu$. Then, for any $a \in (0,1)$

$$\mathbb{P}[Z \geq 1-a] \geq \frac{\mu - (1-a)}{a}$$

We needed to show that

$$\mathbb{P}_{S\sim D^m}\left[L_D(A(S)) \geq \frac{1}{8}\right] \geq \frac{1}{7}$$

([2] in the Thm.)

We write (****)

$$\mathbb{P}_{S\sim D^m}\left[L_D(A(S)) \geq 1-\underbrace{\frac{7}{8}}_{\text{will be our } a\in(0,1)}\right] \geq \frac{\mathbb{E}\left[L_D(A(S))\right]-\left(1-\frac{7}{8}\right)}{7/8}$$

$$\geq \frac{1/8}{7/8}$$

$$= \frac{1}{7}$$

□

no prior knowledge!

Corollary: Let X be an infinite domain and let H be the set of all functions from $X \to \{0,1\}$. Then H is not PAC learnable.

VC-Dimension

(Vapnik & Chervonenkis, 1970)

We have seen that **without** any restrictions on the hyp. class, an adversary can always construct a distribution on which the learning alg. will **perform poorly**, while another alg. succeeds on the same distribution.

Why: We had the "freedom" of choosing a target function (i.e., one of the f_i) from **all** possible functions from $C \to \{0,1\}$.

Let's say we have an hyp. class H now. For PAC learnability, the adversary is now limited to the construction of distributions for which $\exists h^* \in H$ with **zero risk!**

As a starting point, given a finite $C \subset X$, we consider how H behaves on C (note that in the No-Free-Lunch thm., the distributions were concentrated on C).

Def. Restricting H to C

Let H be a hyp. class of functions from $X \to \{0,1\}$ and $C = \{c_1, \ldots, c_m\} \subset X$. The restriction of H to C is

$$H_C = \{(h(c_1), \ldots, h(c_m)) : h \in H\}$$

H_C is a set of functions, each represented by a vector $(0,1,0,0,\ldots 1)$ of length $|C|$. In other words, these are all labelings of C that are possible using the $h \in H$.

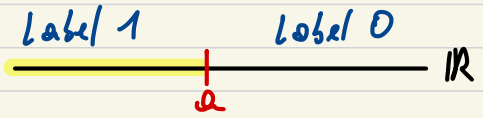
Def. Shattering

For a finite set $C \subset X$, if H_C contains all functions from $C \to \{0,1\}$, we say H shatters C.

EX.: Say $C = \{c_1, c_2\}$. If $H_C = \{(0,0),(0,1),(1,0),(1,1)\}$ then H shatters C, as there are 4 possible labelings of C.

More concrete example: Threshold functions over $\mathbb{R}$.

Visually:



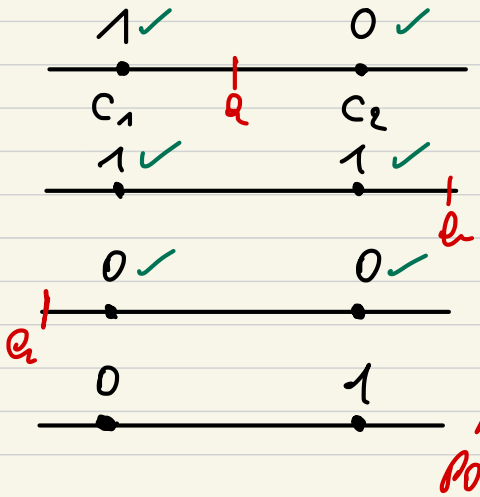
↳ Threshold $\in \mathbb{R}$

[1] Start with $C = \{c_1\}$
$|C| = 1$
If $a = c_1 + 1$, then $h_a(c_1) = 1$
If $a = c_1 - 1$, then $h_a(c_1) = 0$

⇒ as we can choose a freely, we see that we can realize all possible labelings of $C = \{c_1\}$ and thus H shatters C.

[2] We go on with $C = \{c_1, c_2\}$, $|C| = 2$, and let $c_1 \leq c_2$ without loss of generality (w.l.o.g.).

Visual argument:



⇒ H does NOT shatter C! As c_1, c_2 were arbitrary, H does not shatter any set of size 2!

↳ H_C is $\{(1,0), (1,1), (0,0)\}$ ⇒ $|H_C| = 3 \neq 2^{|C|} = 4$!

Corollary (of the No-Free-Lunch thm.) (*)

Let H be a hyp. class of functions from $\mathcal{X} \to \{0,1\}$. Let m be a training set size. Assume there exists a set C of size $2m$ that is shattered by H. Then, for any learning alg. A, there exist a distribution D over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in H$ such that $L_D(h) = 0$, but with probability of at least $1/7$ over the choice $S \sim D^m$, we have $L_D(A(S)) \geq 1/8$.

Remark:

Note that the fact that H shatters C of size $2m$ means that H can realize all labelings of the $2m$ points in C. Hence, we are exactly in the setting of the No-Free-Lunch theorem, as our alg. only receives samples S of size m!

(Consequently, if H is such that it shatters a large set C, then we cannot learn by just seeing half of the samples.)

Def. VC-dimension. The VC-dimension of a hyp. class H, denoted by

$$VC(H)$$

is the maximal size of a set $C \subset X$ that can be shattered by H.

Note: If H shatters sets of arbitrary size, we say that H has infinite VC-dimension ($VC(H) = \infty$)

Corollary: Let H be a hyp class with $VC(H) = \infty$. Then H is not learnable.

Proof: No matter what training set size m we have, there is a set of size $2m$ that is shattered. Hence, by our corollary (*) from the previous page, the claim follows. ∎

(We will see that the converse also holds, i.e., finite VC-dimension implies PAC learnability).

Ex. Threshold functions on $\mathbb{R}$ revisited: We saw that H shatters $C = \{c_1\} \Rightarrow VC(H) \geq 1$. Also, H did NOT shatter sets $C = \{c_1, c_2\}$ of size $2 \Rightarrow VC(H) < 2$.

Thus $VC(H) = 1$.

General strategy to show the VC dim. of H.

$(VC(H) = d)$

$d < \infty$

[1] There exists a set C of size d that is shattered.

[2] Every set of size $d+1$ is not shattered by H.

What about the VC-dim. of finite hyp. classes?

If H is finite, then for any set C, we have

$$|H_C| \leq |H|$$

E.g., if we only have 2 hypotheses and a set C, we can, at most, create 2 different labelings of C. If H shatters C, we know that $|H_C| = 2^{|C|}$ and $|C|$ is the max. size that is shattered, i.e., the VC-dim. We get

$$|H| \geq |H_C| = 2^{|C|} = 2^{VC(H)}$$

$$\Rightarrow \log_2(|H|) \geq VC(H)$$

$$VC(H) \leq \log_2(|H|)$$

Note that this upper-bound can be quite loose!

E.g., k threshold functions $\rightarrow |H| = k$ but $VC(H) = 1$!

Finally, we remark that $VC(H) = d$ means that there exists a set of size d that is shattered. This DOES NOT imply that ALL sets of size d or SMALLER are fully shattered (again, there just exists one of size d that is!)

[I will cover some examples next, before we move on to connect all these things together!]

Plan: Proof of [6] → [1]

Def. (Growth function) Let H be a hypothesis class. Then, the growth function of H, denoted by $\tau_H : \mathbb{N} \to \mathbb{N}$, is defined as

$$\tau_H(m) = \max_{C \subset X : |C| = m} |H_C|$$

This means, $\tau_H(m)$ counts the number of different functions from a set C (of size m) to $\{0,1\}$.

SAUER'S LEMMA: Let H be a hyp. class with $VC(H) \leq d < \infty$. Then, for all m

$$\tau_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

$\Rightarrow$ exponential growth with m!

In particular, if $m > d+1$, we have

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

$\rightarrow$ polynomial growth with m!

Remark: if $VC(H) = d$, then for $m \leq d$, we have $\tau_H(m) = 2^m$.

Why? Because we can shatter a set of size $d \Rightarrow$ all labelings of the points in that set are realizable via functions in H).

Theorem: Let H be a hyp. class and τ_H its growth function. Then, for every distribution D and every $\delta \in (0,1)$, with probability of at least $1-\delta$ (over the choice of S, $S \sim D^m$), we have for a loss function with range $[0,c]$

$$\forall h \in H: |L_D(h) - L_S(h)| \leq c \cdot \sqrt{\frac{8 \cdot \log(\tau_H(2m) \cdot 4/\delta)}{m}}$$

Proof: Recall that

$$|a-b| \geq |a| - |b| \quad (*) \quad \text{(Triangle inequality)}$$

and

$$L_D(h) = \mathop{\mathbb{E}}_{S' \sim D^m} [L_{S'}(h)] \quad (**)$$

$$|(L_D(h) - L_S(h)) - (L_D(h) - L_{S'}(h))| \overset{(*)}{\geq} |L_D(h) - L_S(h)| - |L_D(h) - L_{S'}(h)|$$

This means that

$$|L_{S'}(h) - L_S(h)| \geq \underbrace{|L_D(h) - L_S(h)|}_{(a)} - \underbrace{|L_D(h) - L_{S'}(h)|}_{(b)}.$$

if $(a) > \varepsilon$ and $(b) < \frac{\varepsilon}{2}$, we get

$$|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}$$

Write (a) and (b) in terms of indicator functions. In fact, $|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}$ only holds if (a) and (b) are satisfied. Hence,

$$\mathbb{1}_{|L_D(h) - L_S(h)| > \varepsilon} \cdot \underbrace{\mathbb{1}_{|L_D(h) - L_{S'}(h)| < \frac{\varepsilon}{2}}}_{\text{T1}} \leq \underbrace{\mathbb{1}_{|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}}}_{\text{T2}}$$

T1 Take $\mathbb{E}_{S'}[\cdot]$:

$$\mathbb{E}_{S'}\left[\mathbb{1}_{|L_D(h) - L_{S'}(h)| < \frac{\varepsilon}{2}}\right] = \mathbb{P}_{S'}\left[|L_D(h) - L_{S'}(h)| < \frac{\varepsilon}{2}\right]$$

(Hoeffding) $$\geq 1 - 2e^{-2m\left(\frac{\varepsilon}{2}\right)^2 / (b-a)^2}$$ ($(b-a)^2 \to c^2$)

$$= 1 - 2e^{-m\varepsilon^2/2c^2}$$

T2 Take $\mathbb{E}_{S'}[\cdot]$:

$$\mathbb{E}_{S'}\left[\mathbb{1}_{|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}}\right] = \mathbb{P}_{S'}\left[|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}\right]$$

$$\leq \mathbb{P}_{S'}\left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}\right]$$

Assumption: (we will check this later)

$$m \geq 4c^2 \cdot \varepsilon^{-2} \cdot \log(2) \quad (***)$$

Say we are at equality ($=$). Then,

$$\mathop{\mathbb{P}}_{S'}\left[|L_D(h) - L_{S'}(h)| < \frac{\varepsilon}{2}\right] \geq 1 - 2e^{-m\varepsilon^2/2c^2} \quad \text{(now, set } m = \dots (***)\text{)}$$

$$= 1 - 2e^{-(4c^2\varepsilon^{-2}\cdot\log(2))\varepsilon^2/2c^2}$$

$$= 1 - 2 \cdot e^{-2\cdot\log(2)}$$

$$= 1 - 2e^{-\log(2^2)}$$

$$= 1 - 2/2^2 = 1 - \frac{2}{4} = \frac{1}{2}$$

In combination with $\mathbb{1}_{|L_D(h) - L_S(h)| > \varepsilon}$, we get

$$(****) \quad \mathbb{1}_{|L_D(h) - L_S(h)| > \varepsilon} \leq 2 \cdot \mathop{\mathbb{P}}_{S'}\left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}\right]$$

(under the assumption on $m \geq \dots$ from above).

In fact $(****)$ holds for all $h \in H$! Hence, we can write

$$\mathbb{1}_{\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon} \leq 2 \cdot \mathop{\mathbb{P}}_{S'}\left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}\right]$$

Lets take $\mathop{\mathbb{E}}_{S}[\cdot]$ now!

We get

$$\mathbb{P}_S\left[\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon\right] \leq 2 \cdot \mathbb{P}_{S,S'}\left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}\right]$$

we only have empirical risks!

elements in S or S'

TRICK: Note that $\overbrace{z_i, z_i'}$ are drawn IID from D. Hence, we can easily switch them. What effect do we get?

$$\left(\ell(h, z_i') - \ell(h, z_i)\right) \xrightarrow{\text{(upon switch)}} -\left(\ell(h, z_i') - \ell(h, z_i)\right)$$

$\ell(h, z_i')$	$\ell(h, z_i)$	$\left(\ell(h, z_i') - \ell(h, z_i)\right)$
0	0	0
0	1	-1
1	0	1
1	1	0

In $|L_{S'}(h) - L_S(h)|$, we actually have

$$\left|\frac{1}{m} \cdot \sum_{i=1}^{m}\left(\ell(h, z_i') - \ell(h, z_i)\right)\right|$$

Lets use this insight!

$$\mathbb{P}_S\left[\exists h\in H:\ |L_D(h)-L_S(h)|>\varepsilon\right] \le 2\cdot \mathbb{P}_{S,S'}\left[\exists h\in H:\ |L_{S'}(h)-L_S(h)|>\frac{\varepsilon}{2}\right]$$

$$= 2\cdot \mathbb{P}_{S,S'}\left[\exists h\in H:\ \frac{1}{m}\cdot\left|\sum_{i=1}^{m}\left(\ell(h,z_i')-\ell(h,z_i)\right)\right|>\frac{\varepsilon}{2}\right]$$

$$= 2\cdot \mathbb{E}_{S,S'}\left[\mathbb{1}_{\exists h\in H:\ \frac{1}{m}\cdot\left|\sum_{i=1}^{m}\left(\ell(h,z_i')-\ell(h,z_i)\right)\right|>\frac{\varepsilon}{2}}\right]$$

$$= 2\cdot \mathbb{E}_{S,S'}\left[\mathbb{E}_{\sigma}\left[\mathbb{1}_{\exists h\in H:\ \frac{1}{m}\cdot\left|\sum_{i=1}^{m}\sigma_i\left(\ell(h,z_i')-\ell(h,z_i)\right)\right|>\frac{\varepsilon}{2}}\right]\right]$$

We can do this if σ_i is drawn from a uniform distribution on $\{+1,-1\}$. So $\sigma\sim U^m(\{+1,-1\})$.

↳ uniform distr.

Note: The σ_i's are called RADEMACHER variables!

We have now that

$$\ldots \le 2\cdot \mathbb{E}_{S,S'}\left[\mathbb{P}_{\sigma}\left[\exists h\in H:\ \frac{1}{m}\cdot\left|\sum_{i=1}^{m}\sigma_i\cdot\left(\ell(h,z_i')-\ell(h,z_i)\right)\right|>\frac{\varepsilon}{2}\right]\right]$$

As $\sum_{i=1}^{m}\ldots$ goes over $1,\ldots,m$ and $z_i', z_i\sim D$, we see that a $h\in H$ is only evaluated on at most $2m$ points! So, lets combine S and S' into a set C, $|C|\le 2m$ and restrict H to C.

Since C is finite, we have that H_C is also finite!

Lets look at the $\mathbb{P}_\sigma[\dots]$ term.

$$\mathbb{P}_\sigma\left[\exists h \in H : \frac{1}{m}\cdot\left|\sum_{i=1}^{m}\sigma_i\left(\ell(h,z_i)-\ell(h,z_i')\right)\right| > \frac{\varepsilon}{2}\right] \leq$$

by union bound

$$\sum_{h\in H_c}\mathbb{P}_\sigma\left[\left|\sum_{i=1}^{m}\sigma_i\cdot\frac{1}{m}\cdot\underbrace{\left(\ell(h,z_i')-\ell(h,z_i)\right)}_{=:k_i}\right| > \frac{\varepsilon}{2}\right] =$$

$$\sum_{h\in H_c}\underbrace{\mathbb{P}_\sigma\left[\left|\sum_{i=1}^{m}\sigma_i\cdot\frac{1}{m}\cdot k_i\right| > \frac{\varepsilon}{2}\right]}_{\mathbb{E}_\sigma\left[\mathbb{1}_{\left|\sum\sigma_i\frac{1}{m}k_i\right|>\frac{\varepsilon}{2}}\right]}$$

As all σ_i are drawn IID from $\mathcal{U}(\{+1,-1\})$, we know that the expectation is 0!

$$\to \sum_{h\in H_c}\mathbb{E}_\sigma\left[\mathbb{1}_{\left|\sum\sigma_i\frac{1}{m}k_i\right|>\frac{\varepsilon}{2}}\right] = \sum_{h\in H_c}\mathbb{E}_\sigma\left[\mathbb{1}_{\left|\sum\sigma_i\frac{1}{m}k_i - 0\right|>\frac{\varepsilon}{2}}\right]$$

$$= \sum_{h\in H_c}\mathbb{P}_\sigma\left[\left|\sum_{i=1}^{m}\sigma_i\frac{1}{m}k_i - 0\right| > \frac{\varepsilon}{2}\right]$$

(use Hoeffding inequality)

Side result: $\mathbb{P}_\sigma\left[\left|\sum_{i=1}^{m}\sigma_i\frac{1}{m}k_i - 0\right| > \frac{\varepsilon}{2}\right] \leq 2\cdot e^{-2m\left(\frac{\varepsilon}{2}\right)^2/(2c)^2}$

$$= 2\cdot e^{-2m\varepsilon^2/16c^2}$$

$$= 2\cdot e^{-m\varepsilon^2/8c^2}$$

Combined, we get:

$$\mathbb{P}_\sigma\left[\exists h\in H : \frac{1}{m}\cdot\left|\sum_i\sigma_i\left(\ell(h,z_i')-\ell(h,z_i)\right)\right| > \frac{\varepsilon}{2}\right] \leq 2\cdot e^{-m\varepsilon^2/8c^2}\cdot|H_c|$$

Because we have $\sum_{h\in H_c}\dots$

Finally :)

$$\mathop{\mathbb{P}}_{S}\left[\exists h \in H: |L_D(h) - L_S(h)| > \varepsilon\right] \leq 2 \cdot \mathop{\mathbb{E}}_{S,S'}\left[|H_C| \cdot 2 \cdot e^{-m\varepsilon^2/\rho_C^2}\right]$$

$$= 4 \cdot \mathop{\mathbb{E}}_{S,S'}\left[|H_C| \cdot e^{-m\varepsilon^2/\rho_C^2}\right]$$

Now, we replace $|H_C|$ by $\tau_H(2m)$ => we no longer have a dependency on S or S'!

Overall, we get

$$\mathop{\mathbb{P}}_{S}\left[\exists h \in H: |L_D(h) - L_S(h)| > \varepsilon\right] \leq \underbrace{4 \cdot \tau_H(2m) \cdot e^{-m\varepsilon^2/\rho_C^2}}_{< \delta}$$

$$4\tau_H(2m)\, e^{-m\varepsilon^2/\rho_C^2} < \delta$$

$$e^{-m\varepsilon^2/\rho_C^2} < \delta / 4\tau_H(2m) \quad | \log$$

$$-m\varepsilon^2/\rho_C^2 < \log\left(\delta / 4\tau_H(2m)\right)$$

$$m\varepsilon^2 > \rho_C^2 \cdot \log\left(\frac{4\tau_H(2m)}{\delta}\right)$$

$$\varepsilon > c \cdot \sqrt{\frac{8 \cdot \log\left(4\tau_H(2m)/\delta\right)}{m}}$$

↳ This is the ε-term on our theorem!

Let's check the assumption!

Recall: $m \geq 4C^2 e^{-2} \log(2)$ (Set in ε)

$$\Rightarrow m \geq 4\varepsilon^{-2} \cdot \frac{1}{\frac{2}{\varepsilon^2} \cdot \frac{8 \cdot \log(4 \cdot \tau_H(2m)/\delta)}{m}} \cdot \log(2)$$

$$8 \log\left(4 \tau_H(2m)/\delta\right) \geq 4 \cdot \log(2)$$

$$\Rightarrow \log(\dots) \geq \frac{1}{2} \cdot \log(2) \qquad | e^{\dots}$$

$$\Rightarrow 4 \tau_H(2m)/\delta \geq \underbrace{e^{\frac{1}{2} \cdot \log(2)}}_{\sqrt{2}}$$

$$\Rightarrow \frac{1}{\delta} \geq \frac{\sqrt{2}}{4 \tau_H(2m)}$$

$$\Rightarrow \delta \leq 4 \cdot \tau_H(2m) \cdot \frac{1}{\sqrt{2}}$$

$$\Rightarrow \delta \leq 2\sqrt{2} \cdot \tau_H(2m)$$

δ is in $(0,1)$ -> so, the last inequality is ALWAYS satisfied!

We have shown that $P_S\left[\exists h \in H : |L_D(h) - L_S(h)| \geq \varepsilon\right] < \delta$ (for ε as before). This is equivalent to

$$P_S\left[\forall h \in H : |L_D(h) - L_S(h)| \leq \varepsilon\right] \geq 1 - \delta$$

Req

[6] → [1] Tricky!

Our proof will be different to the book, but show the same claim.

Before, we introduce a very important result in learning theory.

Def. growth function

Let H be a hyp. class. Then the growth function of H, denoted by $\tau_H : \mathbb{N} \to \mathbb{N}$, is defined as

$$\tau_H(m) = \max_{C \subset X : |C| = m} |H_C|$$

# of different functions from a set C of size m to $\{0,1\}$

Observation. If $VC(H) = d$, then for $m \leq d$, we have $\tau_H(m) = 2^m$ (as we can shatter a set of size d → all labelings of d points are realizable via functions in H)

Sauer's lemma

Let H be a hyp. class (as before) with $VC(H) \leq d < \infty$.

Then, for all m

$$\tau_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

(Exponential growth with m.)

In particular, if $m > d+1$, we have

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

(Polynomial growth with m.)

While, we will NOT proof Sauer's lemma, we proof [6]→[1] now.

Thm. Let H be a hyp. class and τ_H it's growth function. Then, for every distribution D and every $\delta \in (0,1)$, with prob. of at least $1-\delta$ over the choice of $S \sim D^m$, we have for a loss function in the range $[0,c]$

$$\forall h \in H: \; |L_D(h) - L_S(h)| \leq c \cdot \sqrt{\frac{8 \cdot \log(\tau_H(2m) \cdot 4/\delta)}{m}}$$

Proof (sorry, it's quite long):

Let's start with the triangle inequality for absolute values:

$$|a-b| \geq |a| - |b| \quad (*)$$

Also, recall that

$$L_D(h) = \mathop{\mathbb{E}}_{S' \sim D^m} \left[L_{S'}(h) \right] \quad (**)$$

This allows us to replace $L_D(h)$ in the theorem by $(**)$. Assume we have $S, S' \sim D^m$, i.e., to random samples of size m. Using $(*)$, we have

$$\left| \left(L_D(h) - L_S(h)\right) - \left(L_D(h) - L_{S'}(h)\right) \right| \geq |L_D(h) - L_S(h)| - |L_D(h) - L_{S'}(h)|$$

Re-arranging terms gives

$$|L_{S'}(h) - L_S(h)| \geq \underbrace{|L_D(h) - L_S(h)|}_{(a)} - \underbrace{|L_D(h) - L_{S'}(h)|}_{(b)}$$

if (a) $> \varepsilon$ and (b) $< \frac{\varepsilon}{2}$, we get

$$|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}$$

We could try to write these conditions on (a) and (b) in terms of indicator functions and formulate an inequality: In fact $|L_{S'}(h) - L_S(h)| > \varepsilon/2$ only holds if both (a) and (b) are satisfied. Hence

$$\mathbb{1}_{|L_D(h) - L_S(h)| > \varepsilon} \cdot \underbrace{\mathbb{1}_{|L_D(h) - L_{S'}(h)| < \frac{\varepsilon}{2}}}_{\boxed{T1}} \leq \underbrace{\mathbb{1}_{|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}}}_{\boxed{T2}}$$

T1 if we take $\mathbb{E}_{S'}[\cdot]$, we get

$$\mathbb{E}_{S'}\left[\mathbb{1}_{|L_D(h) - L_{S'}(h)| < \frac{\varepsilon}{2}}\right] = \mathbb{P}_{S'}\left[|L_D(h) - L_{S'}(h)| < \frac{\varepsilon}{2}\right]$$

(by Hoeffding) $$\geq 1 - 2e^{-2m\left(\frac{\varepsilon}{2}\right)^2/(b-a)^2} \qquad L_c - 0$$

$$= 1 - 2e^{-m\varepsilon^2/2c^2}$$

The other term including S' is $\boxed{T2}$. Again $\mathbb{E}_{S'}[\cdot]$ gives

$$\mathbb{E}_{S'}\left[\mathbb{1}_{|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}}\right] = \mathbb{P}_{S'}\left[|L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}\right]$$

$$\leq \mathbb{P}_{S'}\left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \varepsilon/2\right]$$

Assumption: $m \geq 4c^2\varepsilon^{-2} \cdot \log(2)$

Say, we are at equality, i.e., $m = \dots$, then

$$\mathbb{P}_{S'}\left[|L_D(h) - L_{S'}(h)| < \frac{\varepsilon}{2}\right] \geq 1 - 2e^{-m\varepsilon^2/2c^2}$$

$$- \left(4c^2\varepsilon^{-2}\cdot\log(2)\right)\varepsilon^2/2c^2$$

$$= 1 - 2e$$

$$= 1 - 2e^{-2\log(2)}$$

$$= 1 - 2e^{-\log(2^2)}$$

$$= 1 - 2/(2^2) = 1 - \frac{1}{2} = \frac{1}{2}$$

In combination with $\mathbb{1}_{|L_D(h) - L_S(h)| > \varepsilon}$ (which is not affected by $\mathbb{E}_{S'}[\cdot]$), we get

$$\mathbb{1}_{|L_D(h) - L_S(h)| > \varepsilon} \leq 2 \cdot \mathbb{P}_{S'}\left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}\right]$$

(under the assumption on $m \geq \dots$)

Note that the previous inequality holds for all h. Hence, we can write

$$\mathbb{1}_{\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon} \leq 2 \cdot \mathbb{P}_{S'} \left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2} \right]$$

Let's take $\mathbb{E}_S[\cdot]$ now:

$$\mathbb{P}_{S'} \left[\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon \right] \leq 2 \cdot \mathbb{P}_{S, S'} \left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2} \right]$$

Both are empirical sums

Now comes a TRICK:

Note that all z_i, z_i' are drawn IID from D. Hence, we can easily switch them. So, what changes?

$$\left(\ell(h, z_i') - \ell(h, z_i)\right) \rightarrow -\left(\ell(h, z_i') - \ell(h, z_i)\right)$$

Why?

$\ell(h, z_i')$	$\ell(h, z_i)$	$\left(\ell(h, z_i') - \ell(h, z_i)\right)$
0	0	0
0	1	-1
0	0	0
1	0	1

$\Rightarrow$ switching z_i' and z_i $\Rightarrow$ sign change

In $|L_{S'}(h) - L_S(h)|$, we actually have

$$\left| \frac{1}{m} \sum_{i=1}^{m} \underbrace{\left(\ell(h, z_i') - \ell(h, z_i) \right)}_{\text{if we would multiply by } -1 \text{ at a specific } i, \text{ we would effectively switch } z_i' \text{ and } z_i !} \right|$$

Let's use this insight:

$$\mathbb{P}_S\left[\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon\right] \leq 2 \cdot \mathbb{P}_{S,S'}\left[\exists h \in H : |L_{S'}(h) - L_S(h)| > \frac{\varepsilon}{2}\right]$$

$$= 2 \cdot \mathbb{P}_{S,S'}\left[\exists h \in H : \frac{1}{m} \left| \sum_{i=1}^{m} \left(\ell(h, z_i') - \ell(h, z_i) \right) \right| > \frac{\varepsilon}{2}\right]$$

$$= 2 \cdot \mathbb{E}_{S,S'}\left[\mathbb{1}_{\exists h \in H : 1/m \left| \sum_{i=1}^{m} \left(\ell(h, z_i') - \ell(h, z_i) \right) \right| > \frac{\varepsilon}{2}} \right]$$

$$= 2 \cdot \mathbb{E}_{S,S'}\left[\mathbb{E}_{\sigma}\left[\mathbb{1}_{\exists h \in H : \frac{1}{m} \left| \sum_{i=1}^{m} \left(\ell(h, z_i') - \ell(h, z_i) \right) \cdot \sigma_i \right| > \frac{\varepsilon}{2}} \right] \right]$$

We can do this, if σ_i is drawn from a uniform distribution on $\{-1, +1\}$. The σ_i are so called RADEMACHER variables.

$$\Rightarrow \quad 2 \cdot \mathbb{E}_{S,S'}\left[\mathbb{P}_{\sigma}\left[\exists h \in H : \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i \left(\ell(h, z_i') - \ell(h, z_i) \right) \right| > \frac{\varepsilon}{2} \right] \right]$$

Does not look easier, right? :)

BUT, as $\sum_{i=1}^{m} \dots$ goes over $1, \dots, m$ and $z_i, z_i' \sim D$, we know that a $h \in H$ is only evaluated on at most $2m$ points!
In fact, we combine S and S' into C, $|C| \leq 2m$ and restrict H to C, i.e. H_C. Since C is finite, so is H_C. The question is how fast does H_C grow?

As always, we bound the $\exists$ statement with the union bound (U.B.) and get (for the $\mathbb{P}_\sigma[\cdot]$ term)

$$\mathbb{P}_\sigma\left[\exists h \in H : \frac{1}{m} \cdot \left| \sum_{i=1}^{m} \sigma_i \left(\ell(h, z_i) - \ell(h, z_i') \right) \right| > \frac{\varepsilon}{2} \right] \overset{\text{U.B.}}{\leq}$$

$$\sum_{h \in H_C} \mathbb{P}_\sigma\left[\frac{1}{m} \cdot \left| \sum_{i=1}^{m} \sigma_i \left(\ell(h, z_i') - \ell(h, z_i) \right) \right| > \frac{\varepsilon}{2} \right]$$

$$= \sum_{h \in H_C} \mathbb{P}_\sigma\left[\left| \sum_{i=1}^{m} \sigma_i \cdot \frac{1}{m} \cdot \underbrace{\left(\ell(h, z_i') - \ell(h, z_i) \right)}_{=: h_i} \right| > \frac{\varepsilon}{2} \right]$$

$$- \sum_{h \in H_C} \underbrace{\mathbb{P}_\sigma\left[\left| \sum_{i=1}^{m} \sigma_i \cdot \frac{h_i}{m} \right| > \frac{\varepsilon}{2} \right]}_{\mathbb{E}_\sigma\left[\mathbb{1} \left| \sum_{i=1}^{m} \sigma_i \frac{h_i}{m} \right| > \frac{\varepsilon}{2} \right]}$$

We know that the σ_i are drawn IID from a uniform distr. on $\{-1, +1\}$ $\Rightarrow$ expectation is 0!

We could (without changing anything) write

$$\sum_{h \in H_c} \mathbb{E}_\sigma \left[\mathbb{1} \left| \sum_{i=1}^{m} \left(\sigma_i \cdot \frac{k_i}{m} - 0 \right) \right| > \frac{\varepsilon}{2} \right], \text{ or}$$

$$\sum_{h \in H_c} \mathbb{P}_\sigma \left[\left| \sum_{i=1}^{m} \frac{k_i}{m} \cdot \sigma_i - 0 \right| > \frac{\varepsilon}{2} \right]$$

This seems to be a classic case for Hoeffding again, and we see why we introduced the RADEMACHER variables :)

Note that $k_i := \left(\ell(h, z_i') - \ell(h, z_i) \right)$ and we know the loss is in $[0, c]$ $\Rightarrow$ the difference is in $[-c, c]$.

$$\Rightarrow \mathbb{P}_\sigma \left[\left| \sum_{i=1}^{m} \frac{k_i}{m} \sigma_i - 0 \right| > \frac{\varepsilon}{2} \right] = \mathbb{P}_\sigma \left[\left| \frac{1}{m} \cdot \sum_{i=1}^{m} k_i - 0 \right| > \frac{\varepsilon}{2} \right]$$

$$\leq 2 \cdot e^{-2m \cdot \left(\frac{\varepsilon}{2}\right)^2 / (2c)^2}$$

$$= 2 \cdot e^{-2m\varepsilon^2 / 16c^2} = 2 \cdot e^{-m\varepsilon^2/(8c^2)}$$

Combined with $\sum_{h \in H_c}$, we get

$$\mathbb{P}_\sigma \left[\exists h \in H_c : \frac{1}{m} \cdot \left| \sum_i \left(\ell(h, z_i') - \ell(h, z_i) \right) \cdot \sigma_i \right| > \frac{\varepsilon}{2} \right] \leq |H_c| \cdot 2e^{-m\varepsilon^2/8c^2}$$

FINALLY :), we are ready to conclude that

$$\mathbb{P}_S\left[\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon\right] \le 2 \cdot \mathbb{E}_{S,S'}\left[|H_C| \cdot 2e^{-m\varepsilon^2/8c^2}\right]$$

$$= 4 \mathbb{E}_{S,S}\left[|H_C| \cdot 2e^{-m\varepsilon^2/8c^2}\right]$$

We are now ready to replace $|H_C|$ by $\tau_H(2m)$, remembering that $|C| \le 2m$. We get (as $\tau_H(2m)$ is independent of S, S'):

$$\mathbb{P}_S\left[\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon\right] \le \underbrace{4\tau_H(2m) \cdot e^{-m\varepsilon^2/8c^2}}_{< \delta}$$

$$4\tau_H(2m) \cdot e^{-m\varepsilon^2/8c^2} < \delta$$

$$e^{-m\varepsilon^2/8c^2} < \delta / 4\tau_H(2m) \quad / \log.$$

$$-m\varepsilon^2/8c^2 < \log\left(\delta / 4\tau_H(2m)\right)$$

$$m\varepsilon^2 > 8c^2 \cdot \log\left(4\tau_H(2m)/\delta\right)$$

$$\varepsilon > c\sqrt{\frac{8 \cdot \log\left(4\tau_H(2m)/\delta\right)}{m}}$$

This is our ε from the theorem's claim!

It remains to check if our assumption on $m \geq \dots$ was valid. In particular, this was

$$m \geq 4c^2\tilde{\varepsilon}^2 \log(2)$$

$$m \geq 4c^2 \cdot \frac{1}{c^2 \cdot \frac{8 \ln(4\tau_H(2m)/\delta)}{m}} \cdot \log(2)$$

$$m \geq \frac{4 \cdot m \log(2)}{8 \log(4\tau_H(2m)/\delta)}$$

$$8 \log(4\tau_H(2m)/\delta) \geq 4 \cdot \log(2)$$

$$\log(\;) \geq \frac{1}{2} \cdot \log(2)$$

$$4\tau_H(2m)/\delta \geq e^{\frac{1}{2}\log(2)} = \sqrt{2}$$

$$\frac{1}{\delta} \geq \sqrt{2}/4\tau_H(2m)$$

$$\delta \leq \frac{4\tau_H(2m)}{\sqrt{2}}$$

$$\delta \leq 2 \cdot \sqrt{2} \cdot \tau_H(2m)$$

As $\delta \in (0,1)$, this last inequality is ALWAYS satisfied $\Rightarrow$ Ok!

Overall, we have shown that

$$\mathbb{P}_S\left[\exists h \in H: |L_D(h) - L_S(h)| > \varepsilon\right] < \delta$$

→ we made this explicit

which is equivalent to

$$\mathbb{P}_S\left[\forall h \in H: |L_D(h) - L_S(h)| \leq \varepsilon\right] \geq 1 - \delta$$

or, in other words "that with prob. of at least $1-\delta$

$$\forall h \in H: \; |L_D(h) - L_S(h)| \leq c \cdot \sqrt{\frac{\delta \cdot \log(4 \cdot \Pi_H(2m)/\delta)}{m}}$$

In the book, it is shown that the supremum over H is bounded, but since, in our case, we have $\forall h \in H$, this holds for the supremum as well!

This establishes the last equivalence [5] → [1] in the fundamental theorem of PAC learning!

LINEAR PREDICTORS

Def. We define the class of **affine functions**

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

with

$$h_{w,b}(x) = \langle w, x \rangle + b = \left(\sum_{i=1}^{d} w_i x_i \right) + b$$

$(x \in \mathbb{R}^d)$

There are different hypothesis classes of **linear predictors** and each class is a composition of a function

$$\phi : \mathbb{R} \to Y$$

with L_d. Example: $\phi = \text{sign}$. In that case, i.e., $\phi = \text{sign}$, we get the hypothesis class of **HALFSPACES**.

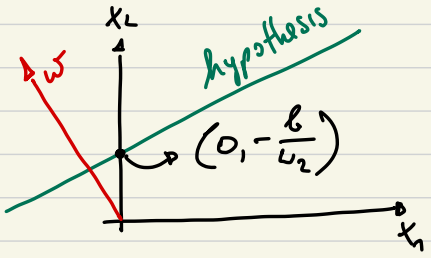
Notation: We will also write L_d as $\{ x \mapsto \langle w, x \rangle + b : w \in \mathbb{R}^d, b \in \mathbb{R} \}$

HALFSPACES Let $X = \mathbb{R}^d$ and $Y = \{+1, -1\}$. Then

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

is called the class of **halfspace hypotheses**.

Example in $\mathbb{R}^2$:



$w = (w_1, w_2)^T$

$\langle w, x \rangle + b = 0$

$w_1 \cdot x_1 + w_2 \cdot x_2 + b = 0 \quad | \; x_1 = 0$

$w_2 x_2 + b = 0$

$\Rightarrow x_2 = -\frac{b}{w_2}$

Remark: Each affine function in $\mathbb{R}^d$ can be re-written as a homogenous linear function in $\mathbb{R}^{d+1}$.

Ex.: Aff. function on $\mathbb{R}^d$: $\langle w, x \rangle + b$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$

Re-written as hom. lin. function on $\mathbb{R}^{d+1}$:

$w' = (b, w_1 \dots w_d)^T$ and

$x' = (1, x_1 \dots x_d)^T$

$\Rightarrow \langle w, x \rangle + b = \langle w', x' \rangle$

Question: What is the VC-dimension of halfspaces?

Theorem: The VC-dimension, $VC(H)$, of the class of homogenous halfspaces, i.e., $H = \{x \mapsto \text{sign}(\langle w, x \rangle) : w \in \mathbb{R}^d\}$ is d!

Proof:

Part [1]: Lower-bound $(VC(H) \geq d)$

Let $C = \{x_1, \dots, x_d\}$ with $x_i \in \mathbb{R}^d$ and

$$x_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}, \dots \quad x_d = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

If we set $w = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}$ for any given labeling $(y_1 \dots y_d)$ of our d data points

we get

$$\langle w, x_i \rangle = \sum_{j=1}^{d} w_j \cdot x_{ij} = y_i \cdot 1 = y_i$$

$\hookrightarrow x_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ — i-th position

We conclude PART [1], i.e., $VC(H) \geq d$

PART [2]: Upper-bound $(VC(H) < d+1)$

We will establish $VC(H) < d+1$ via contradiction.

Assume $C = \{x_1, \dots, x_{d+1}\}$ is shattered by $H \Rightarrow$ All 2^{d+1} labelings can be realized! That means that we have weights w_k, $k = 1, \dots, 2^{d+1}$ with which it is possible to realize the 2^{d+1} labelings.

Lets write:

$$\begin{matrix} d+1 \\ \downarrow \end{matrix} \begin{pmatrix} w_1^T x_1 & w_2^T x_1 & \cdots & w_{2^{d+1}}^T x_1 \\ w_1^T x_2 & w_2^T x_2 & & \vdots \\ \vdots & \vdots & & \\ w_1^T x_{d+1} & w_2^T x_{d+1} & \cdots & w_{2^{d+1}}^T x_{d+1} \end{pmatrix} = X \cdot W$$

$\rightarrow 2^{d+1}$

$\underline{X}$ here is

$$X = \begin{pmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_{d+1} & - \end{pmatrix} \begin{matrix} \downarrow \\ d+1 \end{matrix}$$

$\rightarrow d$

$\underline{W}$ here is:

$$\begin{pmatrix} | & | & & | \\ w_1 & w_2 & \cdots & w_{2^{d+1}} \\ | & | & & | \end{pmatrix} \begin{matrix} \downarrow \\ d \end{matrix}$$

$\rightarrow 2^{d+1}$

$X \cdot W$

$(d+1) \times d \quad \rightarrow \quad d \times 2^{d+1}$

Let $M = X \cdot W$. We see that $\text{sign}(M)$ gives All label combinations.

Claim: The rows of M are linearly independent. Why?

(Reminder: A sequence $v_1, \dots v_k$ in a vector space is lin. independent if

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

is only satisfied for $a_i = 0, i = 1, \dots, k$)

We see that the claim is true as there is no vector a with $a \neq 0$ such that $a^T M = 0$.

$$(- a^T -) \cdot M = u \in \mathbb{R}^{1 \times (2^{d+1})}$$

$1 \times (d+1)$ $\quad$ $(d+1) \times (2^{d+1})$

The k-th element in u is of the form $\boxed{a^T X w_k}$.

As we assume shattering, there is a k where

$$\text{sign}(a) = \text{sign}(X w_k)$$

As a result $a^T X w_k$ is a sum of positive numbers

$\Rightarrow$ this cannot be 0!

Lin. independence of the $d+1$ rows implies $\text{rank}(M) = d+1$.

We also know that

$$M = X \cdot W$$

and

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

(a) only has d-columns

$\Rightarrow$ CONTRADICTION! Our assumption that H shatters $d+1$ points does not hold!

This concludes PART 2, i.e., $VC(H) < d+1$

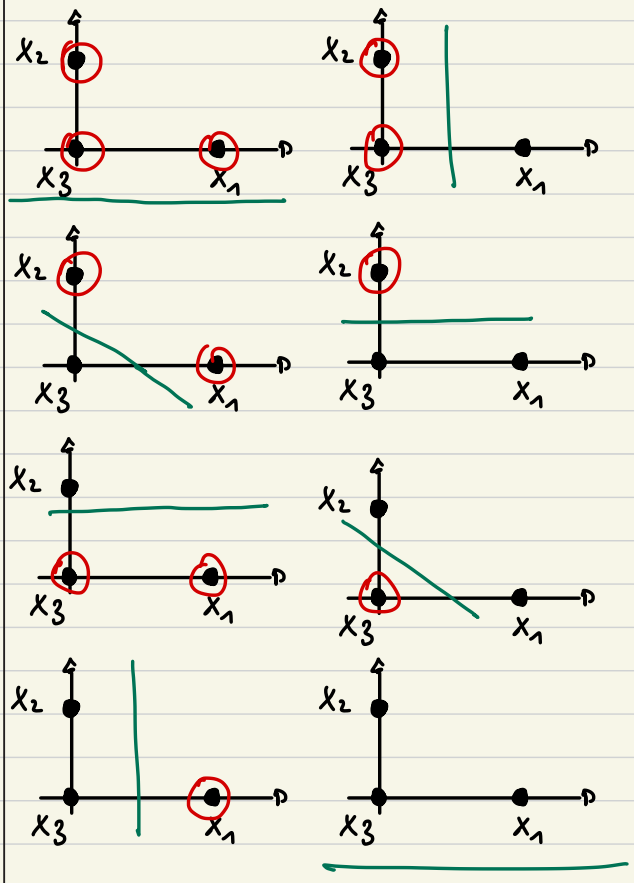
Overall, combining $VC(H) \geq d$ and $VC(H) < d+1$, we get

$$VC(H) = d$$

∎

$\langle w, x \rangle + b = \langle w', x' \rangle \quad x', w' \in \mathbb{R}^3$

Example of halfspaces in $\mathbb{R}^2$ (VC-dim. lower bound)



x_1	x_2	x_3	
1	1	1	✓
1	1	-1	✓
1	-1	1	✓
1	-1	-1	✓
-1	1	1	✓
-1	1	-1	✓
-1	-1	1	
-1	-1	-1	

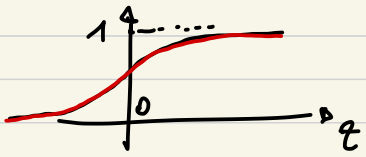
→ VC-dim ≥ 3

Linear predictors - Logistic Regression

Idea: We learn a class of functions from $\mathbb{R}^d \to [0,1]$, i.e., $x \mapsto h(x) \in [0,1]$ and interpret the output as the probability that x has label 1.

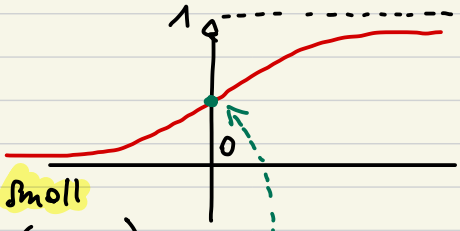
Formally, we compose L_d with ϕ_{sig}, defined as

$$\phi_{sig}(z) = \frac{1}{1+e^{-z}} \quad \text{(Sigmoid function)}$$



$$\Rightarrow \underbrace{H_{sig}}_{\text{Hyp. class}} = \phi_{sig} \circ L_d = \{x \mapsto \phi_{sig}(\langle w, x \rangle) : w \in \mathbb{R}^d\}$$

Interpretation:



$\langle w,x \rangle$ small $\Rightarrow \phi_{sig}(\langle w,x \rangle)$ close to 0

$\langle w,x \rangle$ large $\Rightarrow \phi_{sig}(\langle w,x \rangle)$ close to 1.

Also, when $\langle w,x \rangle = 0 \Rightarrow \phi_{sig}(\langle w,x \rangle) = \frac{1}{2}$

Note that a halfspace hypothesis would always output 0 or 1 no matter if $\langle w,x \rangle$ is large/small or close to zero! (since we have $\text{sign}(\langle w,x \rangle)$)

Since $h(x) \in [0,1]$, we need to specify the "quality" of $h(x)$ given a training set $(x_1, y_1), \dots, (x_m, y_m)$ with, say $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ $\Rightarrow$ This leads us to the definition of a suitable loss function.

Intuition: if $y = +1$, we want $h(x)$ to be large.
if $y = -1$, we want $h(x)$ to be small.
(or, equivalently, $1 - h(x)$ to be large).

We have
$$1 - h(x) = 1 - \frac{1}{1 + e^{-\langle w, x \rangle}}$$
$$= \frac{1 + e^{-\langle w,x \rangle} - 1}{1 + e^{-\langle w,x \rangle}}$$
$$= \frac{e^{-\langle w,x \rangle}}{1 + e^{-\langle w,x \rangle}}$$
$$= \frac{1}{e^{\langle w,x \rangle} \cdot (1 + e^{-\langle w,x \rangle})}$$
$$= \frac{1}{1 + e^{\langle w,x \rangle}}$$

If we write $\boxed{\frac{1}{1 + e^{y \langle w,x \rangle}}}$ (x) and remember that in case of $y=+1$ $h(x)$ should be large and if $y=-1$, then $1-h(x)$ should be large we realize that (x) plays a central role!

Any loss function that would make sense should thus monotonically increase with

$$\frac{1}{1+e^{y\cdot\langle w,x\rangle}}, \text{ or with } \boxed{1+e^{-y\langle w,x\rangle}}$$

In logistic regression we take $\log\left(1+e^{-y\langle w,x\rangle}\right)$ as loss, noting that $\log(\cdot)$ is a monotonic function.

$\Rightarrow$ Given (x_1,y_1) ---- (x_m,y_m), ERM boils down to

$$(*) \quad \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \log\left(1+e^{-y_i\langle w,x_i\rangle}\right)$$

We will later show that this loss is convex in w, which will help solving $(*)$ efficiently.

Next, we look at a probabilistic interpretation of this!

Logistic-Regression – An alternative perspective

The idea is to consider the problem from a statistical perspective. Consider the conditional distribution of a response random variable Y, given an input random variable X, i.e.,

$$\mathbb{P}[Y|X]$$

tells us how precise our predictions are!

Now pick two classes, say 0 and 1 => then Y becomes an indicator variable

$$\mathbb{P}[Y=1] = \mathbb{E}[Y]$$

Also,

$$\mathbb{P}[Y=1|X=x] = \mathbb{E}[Y|X=x]$$

Let's assume now that $\mathbb{P}[Y=1|X=x] = p(x;\theta)$, for some function p that is parametrized by θ. In this case, the conditional likelihood function (under i.i.d. assumption) is

$$\prod_{i=1}^{m} \mathbb{P}[y=y_i|X=x_i] = \prod_{i=1}^{m} p(x_i;\theta)^{y_i} \left(1-p(x_i;\theta)\right)^{1-y_i}$$

as $y_1 \dots y_m$ is a sequence of Bernoulli trials (where each trial has its own success probability – classically, we would have

$$\prod_{i=1}^{m} p^{y_i} \cdot (1-p)^{y_i}$$

Now: $\prod_{i=1}^{m} p_i^{y_i} \cdot (1-p_i)^{1-y_i}$

In the case $\prod_{i=1}^{m} p_i^{y_i} \cdot (1-p_i)^{1-y_i}$

we would get (by max. likelihood) that $\hat{p}_i = 1$ if $y_i = 1$ and $\hat{p}_i = 0$ if $y_i = 0$ $\Rightarrow$ That does not make a lot of sense!

However, the p_i's are not just arbitrary; in fact, they are linked together via $p_i = p(x_i; \theta)$. So p_i must be the same whenever the x_i's are the same and if $p(\cdot;\cdot)$ is a continuous function, similar values of x_i give similar values of p_i.

We will assume that $p(\cdot;\cdot)$ is known, up to the parameters.
$\Rightarrow$ The likelihood function is a function of θ and we can use max. likelihood!

Choices for p: [1] Linear function of x? Not a good choice, as we need p to be in $[0,1]$

[2] let $\log p$ be a linear function of x?
Problem: $\log p$ is unbounded in one direction

↳

13] "logistic transform": let

$$\log\left(\frac{p}{1-p}\right) \text{ a linear function of } x!$$

$$\Rightarrow \text{This means: } \log \frac{p(x;\theta)}{1-p(x;\theta)} = \langle x, \theta \rangle$$

We get:

$$p(x;\theta) = \frac{e^{\langle \theta, x \rangle}}{1 + e^{-\langle \theta, x \rangle}} = \frac{1}{1 + e^{-\langle \theta, x \rangle}}$$

Using this, we can write down the log-likelihood function:

$$\sum_{i=1}^{m} y_i \cdot \log p(x_i;\theta) + (1-y_i) \cdot \log\left(1 - p(x_i;\theta)\right)$$

$$= \sum_{i=1}^{m} \log\left(1 - p(x_i;\theta)\right) + y_i \log p(x_i;\theta) - y_i \log\left(1 - p(x_i;\theta)\right)$$

$$= \sum_{i=1}^{m} \log\left(1 - p(x_i;\theta)\right) + y_i \log\left[\frac{p(x_i;\theta)}{1 - p(x_i;\theta)}\right]$$

$\langle w, x \rangle = \log\left(\frac{p}{1-p}\right)$ (from before)

$$(*) = \sum_{i=1}^{m} \log\left(1 - p(x_i;\theta)\right) + y_i \langle \theta, x_i \rangle$$

We know that $p(x;\theta) = \frac{1}{1+e^{-\langle \theta, x \rangle}} \Rightarrow 1 - p(x;\theta) = \frac{1}{1+e^{\langle \theta, x \rangle}}$

$$(*) = \sum_{i=1}^{m} -\log\left(1 + e^{\langle \theta, x_i \rangle}\right) + \underbrace{y_i \langle \theta, x_i \rangle}_{\log\left(e^{y_i \langle \theta, x_i \rangle}\right)} = \sum_{i=1}^{m} \log\left(\frac{e^{y_i \langle \theta, x_i \rangle}}{1 + e^{\langle \theta, x_i \rangle}}\right)$$

Now, simplifying gives

$$\sum_{i=1}^{m} \log\left[\frac{e^{y_i \langle \theta, x_i \rangle}}{1 + e^{\langle \theta, x_i \rangle}}\right] = \sum_{i=1}^{m} \log\left[\frac{1}{e^{y_i \langle \theta, x_i \rangle} + e^{(1-y_i)\cdot \langle \theta, x_i \rangle}}\right]$$

This would be the objective for max. likelihood. Equivalently, we can minimize the negative log-likelihood!

$$\Rightarrow \operatorname*{argmin}_{\theta} \sum_{i=1}^{m} \log\left(\underbrace{e^{-y_i \langle \theta, x_i \rangle} + e^{(1-y_i)\cdot \langle \theta, x_i \rangle}}_{(*)}\right)$$

For $y_i = 0$: $(*) = 1 + e^{\langle \theta, x_i \rangle}$

For $y_i = 1$: $(*) = e^{-\langle \theta, x \rangle} + 1 = 1 + e^{-\langle \theta, x \rangle}$

We see that this gives the same loss function as before, where we had $Y = \{-1, +1\}$. Here, we have $Y = \{0, 1\}$. But, the probabilistic perspective gives the same loss!

We also see that logistic regression gives a lin. classifier! But, the decision is more "fine-grain", as opposed to using $\operatorname{sign}(\langle w, x \rangle)$ as in case of halfspaces.