Theory HW2 due May 21st.

Lecture 15.

Recap. [Fundamental theorem of statistical learning]

H has a finite VC-dim $\Longrightarrow$ H has the Uniform Convergent Property.

$\Longleftrightarrow$ If $VCdim(H) = d$, then $m_H^{U.C}(\varepsilon, \delta) \le C_2 \cdot \frac{d + \log\frac{1}{\delta}}{\varepsilon^2}$

we prove a weaker version

$$m_H^{U.C}(\varepsilon, \delta) \le C \cdot \frac{d \cdot \log\frac{1}{\delta}}{\delta^2 \varepsilon^2}$$

① Lemma [Sauer-Shelah-Perles] proof in class

$$\tau_H(m) \triangleq \max_{C \subseteq X, |C| = m} |H_C| \le \sum_{i=1}^{d = VCdim(H)} \binom{m}{i} \le O(m^d)$$

② Theorem [Uniform Convergence] proof in book

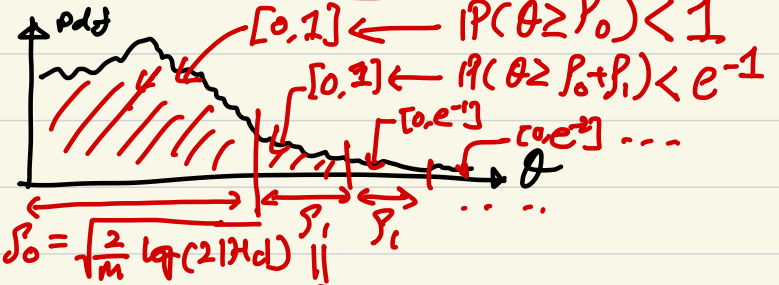
$$|L_D(h) - L_S(h)| \le \frac{4 + \sqrt{\log(\tau_H(2m))}}{\delta \cdot \sqrt{m/2}}$$

* One remaining step. [Lemma A.4 in book]

$$\mathbb{P}\left(\frac{1}{\theta} > \rho\right) \le 2 \cdot |H_C| \cdot e^{-\frac{\rho^2}{2}} \longrightarrow \mathbb{E}[\theta] \le \frac{4 + \sqrt{\log(2|H_C|)}}{\sqrt{m/2}}$$

$\frac{1}{\theta}$: R.V; $e^{-\frac{\rho^2}{2}}$: Sub-Gaussian

Proof sketch> sub-Gaussian pdf



$$\mathbb{E}[\theta] = \int_0^\infty \theta \cdot f(\theta)\, d\theta = \rho_0 + \sum_{i=1}^{\infty} e^{-(i-1)} \cdot (\rho_0 + i \cdot \rho_1) \le \dots$$

$$4 \cdot \sqrt{\tfrac{2}{m}} + \sqrt{\tfrac{2}{m} \log(2|Hd|)}$$

$4 \cdot \rho_1$; $\sqrt{\frac{2}{m} \log(2|Hd|)} = \rho_0$

Chapter 9. Linear Predictors.

- Binary classification with linear classifier: $Y \in \{\pm 1\}$

- Class of affine functions (not linear) $\mathcal{H}_d \triangleq \{ h_{w,b}(x) = \langle w, x \rangle + b : w \in \mathbb{R}^d, b \in \mathbb{R} \}$ (bias; $w^T x$: inner product)

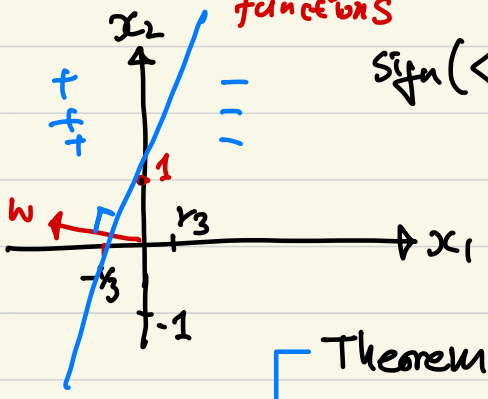
- Prediction at x: $\hat{y} = \text{sign}(\langle w, x \rangle + b)$

- 0-1 loss: $\ell_{w,b}(x, y) = \mathbb{I}(\text{sign}(\langle w, x \rangle + b) \neq y)$

- Learning half spaces:

$$HS_d = \text{sign} \circ \mathcal{H}_d \triangleq \{ \text{sign}(h_{w,b}(x)) = \text{sign}(\langle w, x \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

(HS_d: class of functions; $\circ$: composition of a function $\text{sign}(\cdot)$ and a class.)



$\text{sign}(\langle w, x \rangle + b)$

$w = (-3, 1), b = 1$

decision boundary is $[-3, 1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 1 = 0$

Theorem. For Homogeneous class, i.e. $b = 0$,

$$VCdim(HS_d) = d$$

Proof) ① $\exists C$ with $|C| = d$ s.t. HS_d shatters C.

$C = \{e_1, e_2, \cdots, e_d\}$ is shattered by HS_d.

$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ For any $(y_1, y_2 \ldots, y_d)$, we need to find h_w

s.t. $\text{sign}(\langle e_i, w \rangle) = y_i \quad , \forall i \in [d]$.

let, $w = (y_1, y_2, \cdots, y_d)$, then ⇕

* more generally there is d-degrees of freedom...

② $^{\forall}C$ with $|C| = d+1$, cannot be shattered by HS_d

Consider a set $C = \{x_1, x_2, \cdots, x_d, x_{d+1}\}$ ($x_i \in \mathbb{R}^d$) and labels $y_1, y_2, \cdots, y_d, y_{d+1}$ to be assigned.

proof by contradiction: suppose C can be shattered by HS_d (any y's can be assigned).

$\Rightarrow \exists w \in \mathbb{R}^d$ s.t. $\text{sign}(\langle w, x_1\rangle) = y_1$ for any $(y_1 \cdots y_{d+1})$

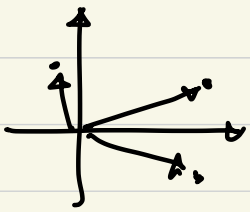
$\text{sign}(\langle w, x_2\rangle) = y_2$

$\vdots$

* Construction of counter example.

for given $C = \{x_1 \cdots x_{d+1}\}$, exists $a = (a_1, a_2, \cdots, a_{d+1}) \in \mathbb{R}$ s.t. $a \neq (0, \cdots, 0)$

$a_1 x_1 + a_2 x_2 + \cdots + a_d x_d + a_{d+1} x_{d+1} = 0$.



because $d+1$ vectors in $\mathbb{R}^d$, are linearly dependent.

from linear Algebra

we let $(y_1 = \text{sign}(a_1), y_2 = \text{sign}(a_2) \cdots) = y$

let $I \triangleq \{i \in [d+1] : a_i > 0\}$

$J \triangleq \{j \in [d+1] : a_j < 0\}$ at least one of them is non-empty set

*Note that $\sum_{i \in I} \underbrace{a_i}_{>0} x_i + \sum_{j \in J} \underbrace{a_j}_{<0} x_j = 0$ by construction

$\rightarrow \sum_{i \in I} a_i x_i = \sum_{j \in J} |a_j| \cdot x_j$

$$0 < \sum_{i \in I} a_i \langle x_i, w\rangle = \left\langle \sum_{i \in I} a_i x_i, w \right\rangle = \left\langle \sum_{j \in J} |a_j| x_j, w \right\rangle = \sum_{j \in J} |a_j| \langle x_j, w\rangle$$

$\text{sign}(a_i) = \text{sign}\langle x_i, w\rangle$ $\quad$ < 0

Theorem 9.3 $VCdim(HS_d)$ with parameter $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ is $d+1$

Proof > (1) $\exists C$. with $|C| = d+1$ s.t. C is shattered by HS_d

$C = \{0, e_1, e_2, \ldots, e_d\}$

label $y_0, y_1, \ldots, y_d$

(2) $\forall C$. with $|C| = d+2$, C is not shattered by HS_d.

same proof with $(w,b) \in \mathbb{R}^{d+1}$ & $d+2$ vectors $C = \{x_1, \ldots, x_{d+2}\}$

---

by Fundamental Theorem of Statistical Learning, we know that [Thm 6.8]

ERM achieves $L_D(h_S) \leq \varepsilon$ w.p $1-\delta$.

(h_S: ERM)

if $m \geq C \cdot \frac{d+\log(1/\delta)}{\varepsilon}$, when D is realizable.

* hard to find ERM when not realizable

Q. How do you find ERM solution halfspaces? (in the realizable case "train data is separable")

$\downarrow$

$\min_{h \in HS_d} L_S(h)$ (L_S: 0-1 loss)

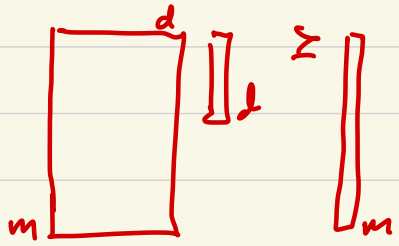
- generally hard
- easy when realizable.

Algorithm 1. Linear Program. ≜ optimization with linear objective & linear constraints

∩ Convex Program

Generic LP: maximize $_{w \in \mathbb{R}^d}$ $\langle u, w \rangle$

subject to $A \cdot w \geq v$: entry-wise inequality



ERM.

find $w \in \mathbb{R}^d$

subject to $\text{sign}(\langle w, x_i \rangle) = y_i$, $(x_i, y_i) \in S$

$\Longleftrightarrow$ $y_i \cdot \langle w, x_i \rangle > 0$: under realizability solution exists.

$\Longleftrightarrow$ $y_i \cdot \langle w, x_i \rangle \geq 1$. : we can always rescale w.

minimize 0

s.t. $y_i \langle w, x_i \rangle \geq 1, \quad \forall i \in [m]$ ⇒ LP-solver ⇒ ERM solution.

Algorithm 2. finds ERM for realizable linear classification.

[Perceptron Algorithm]

Iterative Algorithm:

input: $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$

initialize: $w^{(1)} = (0, \cdots, 0) \in \mathbb{R}^d$

for $t = 1, 2, \cdots$

if $\exists\, i \in [m]$ s.t. $y_i \langle w^{(t)}, x_i \rangle \leq 0$ then

$w^{(t+1)} \leftarrow w^{(t)} + y_i \cdot x_i$

else

output $w^{(t)}$

we want $y_i \langle x_i, w \rangle > 0$

$y_i \cdot \langle w^{(t+1)}, x_i \rangle$

$= y_i \langle w^{(t)} + y_i x_i, x_i \rangle$

$= y_i \langle w^{(t)}, x_i \rangle + y_i^2 \langle x_i, x_i \rangle$

$+1 \cdot \|x_i\|^2$

$> y_i \langle w^{(t)}, x_i \rangle$

Theorem 9.1.

Assume S is separable, $B = \min\{\|w\| : \forall i \in [m],\ y_i \langle w, x_i \rangle \geq 1\}$

$R = \max_i \|x_i\|$.

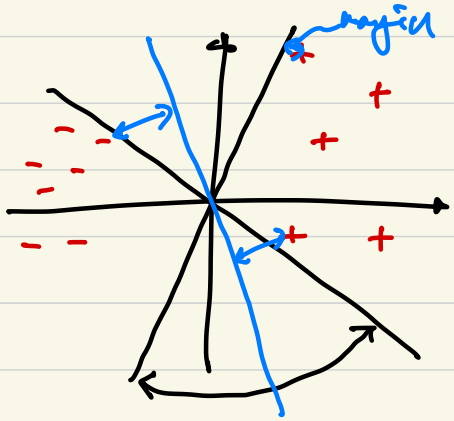
then Perceptron Algorithm stops after at most $(RB)^2$ iterations.

↑ dependence in B is suboptimal!

*when it stops, it finds an ERM with $y_i \langle w^{(t)}, x_i \rangle > 0$, $\forall i \in [m]$

Proof> let $w^* \in \arg\min_w \|w\|$ s.t. $y_i \langle w, x_i \rangle \geq 1$ $\forall i \in [m]$.

↳ max-margin separator



claim: $\frac{\langle w^*, w^{(t+1)} \rangle}{\|w^*\| \cdot \|w^{(t+1)}\|} \geq \frac{\sqrt{T}}{R \cdot B}$

because $L.H.S = \cos(\sphericalangle(w^*, w^{(t+1)})) \leq 1$

↳ $T \leq R^2 B^2$

proof of claim> $\|w^*\| \leq B$, $\langle w^*, w^{(t+1)} \rangle \geq T$ (*), $\|w^{(t+1)}\| \leq R\sqrt{T}$ (**)

(*) $\langle w^*, w^{(t+1)} \rangle - \langle w^*, w^{(t)} \rangle = \langle w^*, w^{(t+1)} - w^{(t)} \rangle$

$= \langle w^*, y_i x_i \rangle$ ← by Algorithm

$= y_i \langle w^*, x_i \rangle \geq 1$

$= 1$ ≥ 1 by def

$\Rightarrow \langle w^*, w^{(t+1)} \rangle = \sum_{t=0}^{T} \{ \langle w^*, w^{(t+1)} \rangle - \langle w^*, w^{(t)} \rangle \} \geq T$

$$(**)\quad \|w^{(t+1)}\|^2 = \|w^{(t)} + y_i x_i\|^2$$

$$= \|w^{(t)}\|^2 + \underbrace{y_i^2}_{=1} \cdot \underbrace{\|x_i\|^2}_{\le R^2} + \underbrace{2 y_i \langle w^{(t)}, x_i \rangle}_{\le 0 \text{ by algorithm}}$$

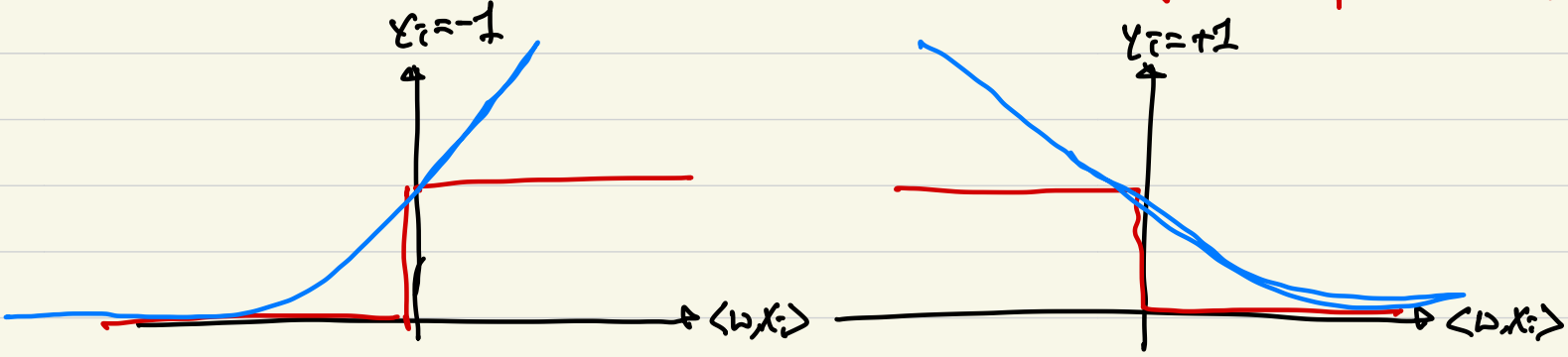
$$= \|w^{(t)}\|^2 + R^2$$

$$\|w^{(T+1)}\|^2 \le T \cdot R^2$$

Algorithm 3. logistic Regression : surrogate loss

True loss: 0-1 loss, $\ell(h_{w,\bullet}, (x,y)) = \mathbb{I}(y \neq \text{sign}(\langle w, x \rangle))$

Surrogate loss: logistic loss, $\ell_w(x,y) = \log(1 + \exp(-y_i \langle w, x_i \rangle))$



* we choose Convex, Smooth surrogate loss

minimize $\frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i \langle w, x_i \rangle))$

Algorithm 4. online Perceptron Algorithm

[much more on online learning at CSE541 Interactive machine learning.]

Data come in an online fashion

time: 1, 2, ..., t, t+1

- receive x_t
- choose h_t, Predict $h_t(x_t)$
- receive y_t
- pay loss: $\ell_{h_t}(x_t, y_t)$

Init $w^{(1)} = 0$ ($\longrightarrow 0$)

for $t = 1, \ldots, t, \ldots T$

- receive x_t
- Predict $p_t = \text{sign}(\langle w^{(t)}, x_t \rangle)$
- If $y_t \langle w^{(t)}, x_t \rangle \le 0$,

$w^{(t+1)} \longleftarrow w^{(t)} + y_t \cdot x_t$

else

$w^{(t+1)} \leftarrow w^{(t)}$

In an online problem, we care about how many mistakes we make over time.

Theorem 11.16 $R = \max_t \|x_t\|$, if realizable, with $\exists w^*$ $y_i \langle w^*, x_i \rangle \geq 1, \forall t$.

$$\#\text{ of mistakes} \leq R^2 \|w^*\|^2$$

proof> [HW3 Problem 5].