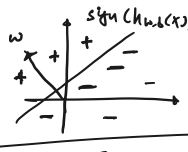
Chapter 9. Linear Predictors.

- class of affine functions $\mathcal{H}_d = \{ h_{w,b}(x) = \langle w, x \rangle + b : w \in \mathbb{R}^d, b \in \mathbb{R} \}$ (bias)
  - not linear.
  - $\langle w, x \rangle$: inner product, $\langle w, x \rangle = w^T x = \sum_{i=1}^{d} w_i x_i$
- Prediction at x : $\hat{y} = \text{sign}(\langle w, x \rangle + b)$
- 0-1 loss : $L_{(x,y)}(h_{w,b}) = \mathbb{1}(\text{sign}(\langle w, x \rangle + b) \neq y)$
- Learning half spaces = linear classifier for binary classification $Y = \{-1, +1\}$

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

(composition of a function and a class.)



Theorem 9.2. For homogeneous class, i.e., $b = 0$, $VCdim(HS_d) = d$.

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

$C = \{e_1, e_2, \dots, e_d\}$ is shattered by HS_d, because for any $(y_1, y_2, \dots, y_d)$, $w = (y_1, y_2, \dots, y_d)$ gives $\text{sign}(\langle w, e_i \rangle) = y_i$

② $\forall C$ with $|C| = d+1$ cannot be shattered by HS_d.

$C = \{x_1, x_2, \dots, x_d, x_{d+1}\}$ (different), suppose C can be shattered by HS_d

Label: $y_1, y_2, \dots y_d, y_{d+1}$; $\text{sign}(\langle w, x_i \rangle) = y_i$

s.t. $\text{sign}(a_1), \text{sign}(a_2), \dots$

& $a_1 \cdot x_1 + a_2 \cdot x_2 + \dots + a_{d+1} x_{d+1} = 0$. ← this exists because $d+1$ vectors in $\mathbb{R}^d$.

a

$I \triangleq \{i : a_i > 0\}$, $J \triangleq \{j : a_j < 0\}$, at least one cannot be empty.

Suppose both nonempty.

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

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

($0 < a_j \cdot \langle x_j, w \rangle$)

- Hence, contradiction.
- If J empty same contradiction with $\sum_{j \in J} |a_j| \langle x_j, w \rangle \geq 0$.

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

① $\exists C$ with $|C| = d+1$ shattered by HS_d.

$C = \{e_0, e_1, \dots, e_d\}$

Label $y_0, y_1, \dots, y_d$ → we let $b = y_0$, $w_i = y_i, \dots, w_d = y_d$. then $\forall x \in C$, $\langle w, x \rangle + b = y_i$

② Same proof in $\mathbb{R}^{d+1}$ and $d+2$ vectors.

$\Rightarrow$ 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 if Not Realizable

*Know Quantitative version of Fundamental Theorem of statistical learning [Thm 6-8]

Q. How do you find ERM solution half spaces? (in the realizable, "separable" case) → $\min_w L_S(w)$

- generally hard
- easy when realizable.

① Linear Program : optimization with linear objective and linear constraints.

(⇑ Convex optimization)

$$\max_{w \in \mathbb{R}^d} \langle u, w \rangle$$

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

$$\begin{bmatrix} & A & \\ & (m \times d) & \end{bmatrix} \begin{bmatrix} w \\ (d \times 1) \end{bmatrix} \geq \begin{bmatrix} v \\ (m \times 1) \end{bmatrix}$$

ERM ← L.P. solvers.

"find w s.t. $\operatorname{sign}(\langle w, x_i\rangle) = y_i$"

$\Leftrightarrow$ $y_i \cdot \langle w, x_i\rangle > 0$. (such w exists under realizability)

claim $\Leftrightarrow$ $y_i \langle \tilde{w}, x_i\rangle \geq 1$, because we can use $\tilde{w} = c \cdot w$ instead with arbitrary c.

min 0
s.t. $y_i \langle w, x_i\rangle \geq 1$, $\forall i \in [m]$

$y_i \cdot x_i^T$ — m — d — $\tilde{w}$ — $\geq$ — 1

---

② Another algorithm that finds an ERM for realizable linear classification.

Iterative algorithm:

input: $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$

initialize: $w^{(1)} = (0, \dots, 0)$

for $t = 1, 2, \dots$

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

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

else

output $w^{(t)}$.

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

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

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

$= y_i \langle w^{(t)}, x_i\rangle + y_i^2 \|x_i\|^2$ (if $\|x_i\| = 1$)

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

Theorem 9.1

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

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

[separable in B is subspace]

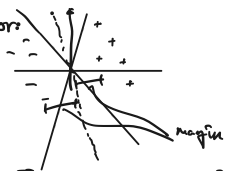
then Perceptron Algorithm stops after at most $(RB)^2$ iterations.

- when it stops it finds an ERM with $y_i \langle w^{(t)}, x_i\rangle > 0$, $\forall i \in [m]$
- Perceptron is PAC learner with $m_H(\varepsilon, \delta) \geq \frac{\log \frac{d}{\delta}}{\varepsilon}$

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: $\frac{\langle w^*, w^{(T+1)}\rangle}{\|w^*\| \cdot \|w^{(T+1)}\|} \geq \frac{\sqrt{T}}{RB} \Rightarrow T \leq (RB)^2$

$1 \geq \cos\angle(w^*, w^{(T+1)}) =$

↑ $\|w^*\| = B$ (def.), $\langle w^*, w^{(T+1)}\rangle \geq T$ (*), $\|w^{(T+1)}\| \leq R\sqrt{T}$ (**)

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

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

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

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

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

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

$= \|w^{(t)}\|^2 + \underbrace{y_i^2}_{1} \underbrace{\|x_i\|^2}_{\leq R} + \underbrace{2 y_i \langle w^{(t)}, x_i\rangle}_{\leq 0}$

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

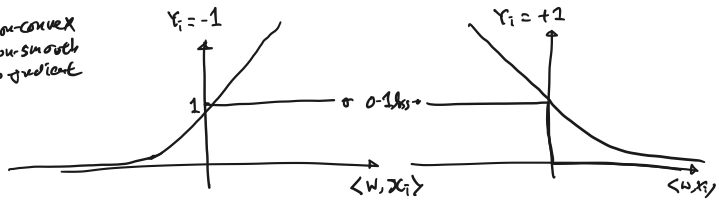
$\|w^{(T+1)}\|^2 \leq T \cdot R^2$

---

③ logistic regression: surrogate loss

true loss: 0-1 loss; $l_{(x,y)}(w) = \mathbb{I}(\operatorname{sign}(\langle w, x\rangle) \neq y)$

- non-convex
- non-smooth
- no gradient



surrogate $l: \mathbb{R} \times Y \to \mathbb{R}_+$ s.t. $l(w, (x, y)) = l(\langle w, x\rangle, y)$

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

- convex
- smooth

ERM: $\min_{w \in \mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp(-y_i\langle w,x_i\rangle))$.

---

*Online Perceptron Algorithm [much more on online learning at CSE 541. Interactive Machine Learning]

Data come in an online fashion.

$1, 2, \ldots, t, t+1, \ldots$

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

Initialize $w^{(1)} = (0, 0, \ldots, 0)$

for $t = 1 \ldots T$

- receive x_t
- Predict $p_t = \text{sign}(\langle w^{(t)}, x_t\rangle)$
- If $y_t\langle w^{(t)}, x_t\rangle \le 0$
  - $w^{(t+1)} \leftarrow w^{(t)} + y_t x_t$
- else $w^{(t+1)} \leftarrow w^{(t)}$

We care about, in an online learning problem, how many mistakes you make.

Theorem 21.16. $R = \max_t \|x_t\|$, if realizable, i.e. $\exists w^*$ s.t. $y_t\langle w^*, x_t\rangle \ge 1 \ \forall t$,

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

proof > [HW1 Prob 6.]