Machine

ERINNERUNG:

- $X \subseteq \mathbb{R}^n$ SAMPLES, $Y = \mathbb{R}, \{0, \dots, k\}$ labels

  $Z = X \times Y$

- D WAHRSCHEINLICHKEITS VERTEILUNG AUF $X \times Y$.

- $\mathcal{H}$ HYPOTHESIS SPACE, RAUM DER ZU BETRACHTENDEN FUNKTIONEN.

- $S \in Z^m$ TRAININGSDATEN

- $\ell : \mathcal{H} \times Z \longrightarrow \mathbb{R}_+$ LOSS FUNKTION

  WOBEI $\ell(h, -) : Z \to \mathbb{R}_+$ MESSBAR / INTEGRIERBAR $\forall h \in \mathcal{H}$.

  Bsp.:
  - 0-1 loss $\ell_{0-1}(h, (x,y)) = \begin{cases} 1 & h(x) \neq y \\ 0 & \text{SONST} \end{cases}$
  - SQUARE loss $\ell_{sq}(h, (x,y)) = (h(x) - y)^2$

- $L_D : \mathcal{H} \longrightarrow \mathbb{R}_+$ RISK FUNKTION

  $h \longmapsto \mathop{E}_{z \sim D}(\ell(h, z))$

- $L_-(-) : Z^m \times \mathcal{H} \longrightarrow \mathbb{R}_+$ EMPIRICAL RISK.

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

  WOBEI $S = (z_1, \dots, z_m)$

• i.i.d. BETRACHTEN PRODUKTMASS AUF Z^m

EMPIRICAL RISK MINIMIZATION

~~FINDE $h \in \mathcal{H}$ S.D. L~~

SEI $S \in Z^m$. FINDE $h \in \mathcal{H}$ S.D. $L_S(h)$ MINIMIERT WIRD.

WIR UNTERSUCHEN, UNTER WELCHEN BEDINGUNGEN ERM ZU EINER GUTEN HYPOTHESE FÜHRT, ALSO $L_D(h)$ KLEIN

DAZU:

- REALIZABILITY: $\exists h^* \in \mathcal{H}$, S.D. $L_D(h^*) = 0$

  $\Rightarrow L_S(h^*) = 0$ F.S.

  $\Rightarrow$ ~~~~ FÜR $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$ GILT $L_S(h_S) = 0$ F.S.

- ES GIBT EINEN LERN ALGORITHMUS, DER h_S BESTIMMT.
- BETRACHTEN IM FOLGENDEN KLASSIFIKATION $\mathcal{Y} = \{0,1\}$ UND $\ell = \ell_{0-1}$

SATZ: SEI $|\mathcal{H}| < \infty$, $\varepsilon, \delta \in (0,1)$, $m \geq \frac{1}{\varepsilon} \log\left(\frac{|\mathcal{H}|}{\delta}\right)$

SEI D W-VERT AUF Z, S.D. REALIZABILITY ERFÜLLT IST FÜR $\mathcal{H}$.

DANN GILT:

MIT W'KT $\geq 1 - \delta$ BZGL $S \in Z^m$ IST

$$L_D(h_S) \leq \varepsilon \quad \text{FÜR JEDES } h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$$

MIT ANDEREN WORTEN:

FÜR m GROSS GENUG FÜHRT ERM ZU EINER HYPOTHESE, DEREN OBERE FEHLER SCHRANKE GEGEBEN IST DURCH ε.

$$L_D(h_S) \leq \varepsilon.$$

BEWEIS: BETRACHTEN DIE ZV. UND $h_S \in \operatorname{argmin}_{h \in H} L_S(h)$

$$\begin{aligned} Z^m &\longrightarrow \mathbb{R}_+ \\ S &\longmapsto L_D(h_S) \end{aligned}$$

WOLLEN DAS W-MASS DER MENGE $\{ S \in Z^m \mid L_D(h_S) \leq \varepsilon \}$ ABSCHÄTZEN.

SEI $H_B = \{ h \in H \mid L_D(h) > \varepsilon \}$ "SCHLECHTE" HYPOTHESEN

$M = \{ S \in Z^m \mid \exists h \in H_B . L_S(h) = 0 \}$ "SCHLECHTE" TRAININGSDATEN.

WEGEN REALIZABILITY GILT:

$$L_S(h_S) = 0 \Rightarrow L_D(h_S) > \varepsilon \Leftarrow \exists h \in H_B \text{ mit } L_S(h) = 0$$

ALSO IST $\{ S \in Z^m \mid L_D(h_S) > \varepsilon \} \subseteq M$

WEITER $M = \bigcup_{h \in H_B} \{ S \in Z^m \mid L_S(h) \overset{=0}{\cancel{\leq \varepsilon}} \}$

Also:

$$D^m(L_D(h_S) > \varepsilon) \leq D^m(M) \leq \sum_{|\mathcal{H}_B|} D^m(L_S(h) = 0) \qquad (*)$$

Sei $h \in \mathcal{H}_B$ fest.

Jetzt gilt $L_S(h) = 0 \Leftrightarrow \ell(h, z_i) = 0 \quad \forall i = 1, \dots, m \qquad (S = (z_1, \dots, z_m))$

Somit:

$$D^m(L_S(h) = 0) = D^m(\{ S \in Z^m \mid \ell(h, z_i) = 0 \ \forall i \}) \overset{\text{i.i.d.}}{=} \prod^m D(\ell(h, z_i) = 0)$$

Wir haben $D(\ell(h, z_i) \neq 0) = D(\{ (x,y) \mid h(x) \neq y \})$

$$= L_D(h) \quad \left(= \int \mathbb{1}_{h(x) \neq y} \right)$$

Also

$$D(\ell(h, z_i) = 0) = 1 - L_D(h) \leq 1 - \varepsilon$$

Da $h \in \mathcal{H}_B \Rightarrow L_D(h) > \varepsilon$

$$\overset{(*)}{\Longrightarrow} \quad D^m(L_D(h_S) > \varepsilon) \leq |\mathcal{H}_B| (1-\varepsilon)^m \leq |\mathcal{H}| (1-\varepsilon)^m.$$

Wir haben $1 - \varepsilon \leq e^{-\varepsilon}$

$$\Rightarrow \quad D^m(L_D(h_S) > \varepsilon) \leq |\mathcal{H}| e^{-\varepsilon m}$$

FÜR $m \geq \frac{1}{\varepsilon} \log\left(\frac{|H|}{\delta}\right)$ GILT ALSO

$$D^m\left(L_D(h) > \varepsilon\right) \leq \delta \Rightarrow D^m\left(L_D(h) \leq \varepsilon\right) 1-\delta$$

$\square$

PROBABLY APPROXIMATELY CORRECT LEARNING

DEF: (PAC - LEARNABILITY)

EINE ~~HYP.~~ KLASSE $\mathcal{H}$ ~~HE~~ IST PAC - LERNBAR (BZGL. $Z = X \times Y$
ℓ LOSS FCT)

FALLS ES EINE FUNKTION $m_{\mathcal{H}} \cdot (0,1)^2 \to \mathbb{N}$

UND EIN LERN ALGORITHMUS EXISTIERT, S.D. FOLGENDES GILT:

- $\forall \varepsilon, \delta \in (0,1)$
- $\forall$ DISTR. D AUF Z, S.D. $\mathcal{H}$ REALIZABILITY BZGL D ERFÜLLT.

WENN DER ALGORITHMUS MIT i.i.d. TRAININGDATEN DER GRÖßE

$m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ TRAINIERT WIRD,

DANN GIBT DER ALGORITHMUS EINE HYPO h ZURÜCK

S.D. $L_D(h) \leq \varepsilon$ MIT W'KT $\geq 1-\delta$ (IN ~~SAMPLES~~ TRAINING)

BSP: $\mathcal{H}$ ENDL. $\Rightarrow$ ALGO ERM

PARAMETER IN PAC-LEARNING

- ε ACCURACY PARAM. WIE GENAU IST DER LEARNER
- δ CONFIDENCE PARAM: WIE WAHRSCHEINLICH
- $m_{\mathcal{H}}$ "SAMPLE COMPLEXITY".

DEF: ~~DIE~~ ~~DER~~ SAMPLE COMPLEXITY UM MIT $\mathcal{H}$ ZU LERNEN IST DAS MINIMUM DER FKT. ~~S.D.~~ $m_{\mathcal{H}}(\varepsilon, \delta)$ S.D. $\mathcal{H}$ PAC-LERNBAR IST.

KORO: JEDE ENDL. KLASSE $\mathcal{H}$ IST PAC-LERNBAR MIT SAMPLE COMPLEXITY $m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{\varepsilon} \right\rceil$

OHNE REALISIERBARKEIT:

ERINNERUNG: $h^* \in \mathcal{H}$ S.D. $L_{\mathcal{D}}(h) = 0$

FÜR KLASSIFIKATION IMPLIZIERT DIES EINE FKTALE RELATION ZU SAMPLES UND LABELS

DIES MUSS NICHT IMMER GEGEBEN SEIN.

STATTDESSEN SCHÄTZEN WIR MIT DEM MINIMUM AB.

2

DEF: (AGNOSTIC PAC-LEARNING)

$\mathcal{H}$ IST AGNOSTIC PAC-LEARNABLE BZGL. ℓ

FALLS $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ EXISTIERT UND LERN ALGO A S.D.

- $\forall\ (\varepsilon, \delta) \in (0,1)$
- $\forall\ D$

WENN DER ALG. AUF $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d SAMPLES LÄUFT, EINE $h \in \mathcal{H}$ ZURÜCKGIBT S.D.

$$L_D(h) \leq \min_{h' \in \mathcal{H}} L_D(h) + \varepsilon \quad \text{MIT WK'T } \geq 1-\delta$$

BEM.: OFFENBAR VERALLG. DIES PAC-LEARNING

BSP: EINES MIN. ~~PAC~~ FÜR KLASSIFIKATION.

D VERT. AUF $\mathcal{X} \times \{0,1\}$

DANN IST

$$f_D(x) = \begin{cases} 1 & \text{FALLS } P(y=1 \mid x) \geq \frac{1}{2} \\ 0 & \text{SONST} \end{cases}$$

EINE OPTIMALE HYPOTHESE

ÜB: ZEIGE DAS

Wollen jetzt zeigen, dass endl. ~~Klassen~~ Hypo Klassen

agnostisch PAC-lernbar sind.

<u>Idee:</u>

<u>Def:</u> (ε-representative sample)

$S \in Z^m$ Trainingsdaten sind ε-repräsentativ

(bzgl. Z, H, ℓ, D) falls

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

<u>Lemma:</u> S $\frac{\varepsilon}{2}$-rep. dann gilt für $h_S \in \arg\min_{h \in H} L_S(h)$

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

<u>Bew:</u> $L_D(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h) + \frac{\varepsilon}{2} \leq L_D(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$

$\square$

Dies impliziert: ~~ERM~~ damit ERM agn PAC-lernbar ist

reicht es, dass mit Wkt mind. $1-\delta$, S ε-rep. ist

DEF: $\mathcal{H}$ HAT DIE UNIFORM CONVERGENCE EIG. (BZGL $\mathcal{Z}, \ell$)

FALLS $\exists\ m_{\mathcal{H}}^{uc} : (0,1)^2 \to \mathbb{N}$ S.D. $\forall \varepsilon, \delta \in (0,1)$ $\forall D$ auf $\mathcal{Z}$

GILT: FÜR $S \in \mathcal{Z}^m$ MIT $m \geq m_{\mathcal{H}}^{uc}(\varepsilon, \delta)$

DANN IST S MIT WKT $\geq 1-\delta$ ε-REP.

$ERM_{\mathcal{H}}$ IST AGNOSTIC PAC LEARNER FALLS $\mathcal{H}$ UNIFORM CONVERGENCE ERFÜLLT

SATZ: $|\mathcal{H}| < \infty$ DANN IST $\mathcal{H}$ AGN. PAC LEARNER

WIR ZEIGEN $|\mathcal{H}| < \infty \Rightarrow \mathcal{H}$ HAT UC EIG.

IDEE WIE VORHER.

WIR SCHÄTZEN

$$D^m(\{S \mid \exists h \in \mathcal{H} : |L_S(h) - L_D(h)| > \varepsilon\}) \text{ AB}$$

ES GILT

$$D^m(\{S \mid \exists h : |L_S(h) - L_D(h)| > \varepsilon\}) \leq \sum_{h \in \mathcal{H}} D^m(|L_S(h) - L_D(h)| > \varepsilon)$$

INTUITION:

$\lim P(\frac{1}{m} \sum (x_i - E(x_i))) = 0$

PROBLEM: NUR ASYMPTOTISCHES VERHALTEN

LEMMA: (HOEFFDING UNGL.)

$\theta_1, \dots, \theta_m$ i.i.d. ZV. $\forall i \; E(\theta_i) = \mu$

$P(a \leq \theta_i \leq b) = 1$ DANN GILT $\forall \varepsilon > 0$

$P(|\frac{1}{m} \sum^m \theta_i - \mu| > \varepsilon) \leq 2e^{-2m\varepsilon^2/(b-a)^2}$

DAMIT KÖNNEN WIR JETZT ABSCHÄTZEN (*) $\theta_i = \ell(h, z_i)$ ANN $\ell \in [0,1]$

$D^m(\{S \mid \exists h \in H, |L_S(h) - L_D(h)| > \varepsilon\})$

$\leq \sum^m D^m(|L_S(h) - L_D(h)| > \varepsilon) \leq 2|H| e^{-2m\varepsilon^2}$

$\Rightarrow m \geq \frac{\log(2|H|/\delta)}{2\varepsilon^2} \Rightarrow D^m(\ast) \leq \delta$

11

WIR HABEN GEZEIGT:

SEI $\mathcal{H}$ ENDL., $\ell : \mathcal{H} \times Z \to [0,1]$ ZUGEH. FKT.

DANN ERFÜLLT $\mathcal{H}$ DIE UK EIG.

MIT SAMPLE COMPLEX $m_{\mathcal{H}}^{uc}(\varepsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil$

WEITERHIN GILT: $\mathcal{H}$ IST AGN. PAC. LERNBAR

MIT ERM ALGO. MIT SAMPLE COMPLEX

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{uc}\left(\frac{\varepsilon}{2}, \delta\right) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil$$