LECTURE 3:

RECALL AGNOSTIC PAC - LEARNING:

EINE HYPOTHESEN KLASSE/RAUM $\mathcal{H}$ IST AGNOSTISCH - PAC - LERNBAR

BZGL Z - UND LOSS $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ FALLS EINE FKT

$m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ UND EIN LERN ALGORITHMUS EX. S.D.

- $\forall \varepsilon, \delta \in (0,1)$
- $\forall$ DISTR. $\mathcal{D}$ AUF Z
- $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$
- $A : Z^m \to \mathcal{Y}^{\mathcal{X}}$

MIT WAHRSCHEINLICHKEIT $\geq 1 - \delta$ $\quad L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$

HABEN GESEHEN: FÜR KLASSIFIKATION, FALLS $\mathcal{D}$ REALISIERBAR IST

SIND ENDL. KLASSEN $\mathcal{H}$ LERNBAR, BAYES OPTIMAL CLASS

$m_{\mathcal{H}}$ IST MONOTON FALLEND

<u>DEF:</u> (ε - REPRÄSENTATIVES SAMPLE.)

$S \in Z^m$ IST ε - REP (BZGL $\mathcal{H}, \ell, \mathcal{D}$) FALLS

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$$

ALSO: S IST ε - REP FALLS FÜR $\mathcal{H}$ DER EMPIRISCHE FEHLER NICHT ZU WEIT WEG

VOM ECHTEN FEHLER IST.

<u>LEMMA:</u> S $\frac{\varepsilon}{2}$ - REP. ; $h_S \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$

$$\Rightarrow L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$$

D.H.: UM ZU ZEIGEN, DASS ERM-RULE AGNOSTISCH PAC-LERNBAR IST

GENÜGT ES ε-REP ZU ZEIGEN MIT WKT $\geq 1-\delta$

DEF: $\mathcal{H}$ HAT UNIFORM CONVERGENCE PROP FALLS

$$m_{\mathcal{H}}^{uc} : (0,1)^2 \longrightarrow \mathbb{N} \quad \text{EX.}$$

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

FALLS $S \in \mathcal{Z}^m$, $m \geq m_{\mathcal{H}}^{uc}(\varepsilon, \delta)$ DANN S HAT ε-REP MIT WKT $\geq 1-\delta$

COR: IF $\mathcal{H}$ HAT UC MIT $m_{\mathcal{H}}^{uc}$ DANN IST $\mathcal{H}$ AGN. PAC-LERNBAR

MIT $m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{uc}(\frac{\varepsilon}{2}, \delta)$

IN DEM FALL IST $ERM_{\mathcal{H}}$ EIN AGN PAC-LERNER

SATZ: ENDL. KLASSEN SIND AGN. PAC-LERNBAR.

BEW: ZEIGEN UC FÜR ENDL. $\mathcal{H}$.

SEIEN $\varepsilon, \delta \in (0,1)$ FIX. WOLLEN ZEIGEN:

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

ES GILT

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

JETZT SCHÄTZEN WIR DIE SUMMANDEN AB

WIR BRAUCHEN EINE QUANTITATIVE VERSION DES GdGZ:

LEMMA: (HOEFFDING'S UNGLEICHUNG)

$\theta_1, \dots, \theta_m$ i.i.d. RV.; $E(\theta_i) = \mu$; $P[a \leq \theta_i \leq b] = 1$ FÜR $a, b \in \mathbb{R}$

DANN GILT $\forall \varepsilon > 0$:

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

NEHMEN JETZT AN $l \in [0,1]$ UND SETZEN $\theta_i = l(h, z_i)$ DANN FOLGT

$$D^m(\{S \mid \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \varepsilon\}) \leq 2|\mathcal{H}| e^{\frac{-2m\varepsilon^2}{(b-a)^2}}$$

DANN GILT FÜR $m > \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}$

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

$\square$

BIAS - COMPLEXITY-TRADEOFF:

IDEE: UM OVERFITTING ZU VERMEIDEN BESCHRÄNKEN WIR $\mathcal{H}$.

$\Rightarrow$ HABEN EINE FORM VON PRIOR KNOWLEDGE!

BSP VON PRIOR KNOWLEDGE:

- WISSEN ÜBER D Z.B. D NORMALVERTEILT MIT PARAM θ
- REALISIERBARKEIT / KLEINES h
- WISSEN KOMPLEXES $\mathcal{H}$ $\rightsquigarrow$ REALISIERBAR ABER OVERFITTING

FRAGE: IST PRIOR KNOWLEDGE NOTWENDIG?

EX. EIN UNIVERSELLER LERNALGO?

SATZ: (NO - FREE - LUNCH)

SEI A LERN ALGO FÜR BIN. KLASS. BZGL 0-1-LOSS.

SEI $m \leq \frac{|\mathcal{X}|}{2}$. DANN EX. D AUF $\mathcal{X} \times \{0,1\}$ S.D.

1. $\exists f: \mathcal{X} \to \{0,1\}$ MIT $L_D(f) = 0$
2. $L_D(A(S)) \geq \cancel{1}/8$ MIT WKT $\geq \frac{1}{7}$.

BEWEIS: $A : Z^m \longrightarrow (Z \to \{0,1\})$

IDEE: SEI $C \subseteq X$ MIT $|C| = 2m$.

SEI $f : C \to \{0,1\}$.

DEF. VERT. D AUF $X \times \{0,1\}$ DIE UNIFORM AUF DEM BILD VON f IST.

ZEIGEN DASS WIR MIT SOLCHEN VERTEILUNGEN $L_D(A(S)) \geq \frac{1}{8}$ ABSCHÄTZEN KÖNNEN.

BETRACHTE $|f_1, \dots, f_T| = |C \to \{0,1\}|$, $T = 2^{2m}$

DEF FÜR f_i: $D_i(x,y) = \begin{cases} \frac{1}{|C|} & f_i(x) = y \\ 0 & \text{SONST} \end{cases}$

WOLLEN ZEIGEN:

$$\max_i \mathbb{E}_{S \sim D_i^m} L_{D_i}(A(S)) \geq \frac{1}{4} \qquad (*)$$

MIT DER MARKOV UNGLEICHUNG

$$P(|X| \geq a) \geq \frac{\mu - a}{1 - a} \qquad a \in (0,1)$$

FOLGT:

$$D^m\left(L_{D_i}(A(S)) \geq \frac{1}{8}\right) \geq \frac{\frac{1}{8}}{\cancel{\frac{3}{4}}\ \frac{7}{8}} = \cancel{\frac{4}{24} = \frac{1}{6}} \ \frac{1}{7}$$

ES BLEIBT $(*)$ ZU ZEIGEN.

WIR BETRACHTEN ~~SA~~ TRAININGSDATEN AUS C.

ES GIBT $k = (2m)^m$ MÖGLICHE SAMPLES AUS C:

$$S_1, \dots, S_k$$

FÜR f_i: SCHREIBE S_j^i FÜR DIE TRAININGSDATEN

$$S_j^i = \left((x_1^j, f_i(x_1^j)), \dots, x_m^j, f_i(x_m^j) \right)$$

WEGEN GLEICHVERTEILUNG AUF ...

HABEN ALLE S_j^i GLEICHE WKT GESAMPLET ZU WERDEN BZGL D_i^m

DAHER GILT $\quad \mathbb{E}_{S \sim D_i^m} \left(L_{D_i}(A(S)) = \frac{1}{k} \sum L_{D_i}(A(S_j^i)) \right)$

JETZT SCHÄTZEN WIR NACH i AB:

$$\max_i \mathbb{E}_{S \sim D_i^m} L_{D_i}(A(S)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i))$$

$$\geq \min_j \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$

WIR SCHÄTZEN JETZT DIE SUMMANDEN ~~AB~~ $L_{D_i}(A(S_j^i))$ AB.

<u>DAZU:</u> $\quad S_j = (x_1^j, \dots, x_m^j) = (x_1, \dots, x_m)$

SEIEN $v_1, \dots, v_p \in C$ S.D. $v_i \neq x_j$ $\quad \forall i = 1, \dots, p$
$\quad j = 1, \dots, m$

$\Rightarrow \quad p \geq m \quad$ DA $\quad |C| = 2m$

JETZT GILT FÜR $\quad h : C \to \{0,1\}$:

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

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