ג' - 19/5/20

הרצאה #5

נמשיך לדבר על מחלקות למידות PAC:

מחלקות היפותזות סופיות

תזכורת:

* ראשית נבחין שמחלקות סופיות לא אומר שהן קטנות הן יכולות להיות עצומות.
* דיברנו על האלגו' הבא עבור מחלקת היפותזות (סופית):

$$ERM_{\mathcal{H}} := A_{ERM}: S \longrightarrow \arg\min_{h\in\mathcal{H}} L_S(h)$$

(מפני ש: $0 \le L_S(h)$ ו- $\mathcal{H}$ מחלקה סופית ⇐ קיים מינימום).

תחת הנחת הריאליזביליות מתקיים: $L_S(h)=0$ (כלומר קיימת $h\in\mathcal{H}$ עבורה נשיג את ההפסד המינימלי!) אז להיפותזה h_S המוחזרת ע"י $ERM_{\mathcal{H}}$ יתקיים $L_S(h_S)=0$ ☺

החוק ERM הוא קונסיסטנטי: מקיים לכל $y_i=h(x_i): i\in[m]$, לכל $S=\{(x_i,y_i)\}_{i\in[m]}$.

* אז מחלקת היפותזות סופית היא למידת-PAC בעזרת כלל ה-ERM, עם סיבוכיות מדגם $m \ge \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$:

המשפט (פורמלית): יהא X מרחב מדגם, $\mathcal{Y}=\{0,1\}$, $\mathcal{H}\subset\mathcal{Y}^{X}$, $|\mathcal{H}|<\infty$. נקבע $0<\varepsilon,\delta<1$. אזי $m \ge \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$ s.t, לכל התפלגות D מעל X ולכל פונק' תיוג f כך שקיים $h^*\in\mathcal{H}$ ← המקיימת את הנחת הריאליזביליות:
המקיים $L_{D,f}(h^*)=0$, בהסתברות לפחות $1-\delta$ (על בחירת S בגודל m) מתקיים: $L_{D,f}(ERM_{\mathcal{H}}(S)) \le \varepsilon$.

הוכחה:

יהא מדגם S. נסמן את $\arg\min_{h\in\mathcal{H}} L_S(h) = ERM_{\mathcal{H}}(S)$.

⊛ מצב שלנו: בוחרים את m, ובוחרים את המדגם $S \mapsto h_S\in ERM_{\mathcal{H}}(S)$ אסטרטגיה שלנו. (נבחין שע"י בחירת m נקבע לנו ε,δ).

⊛ מצב הטבע: בוחר התפלגות D על X, ופונק' תיוג $f: X\to Y$.

⊛ נסמן: $\mathcal{H}_B = \{h\in\mathcal{H} \mid L_{D,f}(h)>\varepsilon\}$ ← מחלקת ההיפותזות ה"רעות" שה-loss שלהן $> \varepsilon$, (האמיתי).

⊛ יהא $S=\{(x_i,y_i)\}_{i\in[m]}$ שנדגמה בהתפלגות $D^m(S)$ (הכוונה רק לקוא' הראשונה - לאיקסים)

נרצה להראות ש: $D^m(\{S \mid \exists h\in ERM_{\mathcal{H}}(S) \text{ s.t: } h\in\mathcal{H}_B\}) \le \delta$

(כלומר - הסיכוי לשלוף היפותזה רעה עבור S $\le \delta$).

מהנחת הריאליזביליות, לכל $h_S \in ERM_{\mathcal{H}}(S)$ מתקיים: $L_{D,f}(h_S)=0$.

נסמן: $M=\{S \mid \exists h\in \mathcal{H}_B,\ L_S(h)=0\}$ ← קב' המדגמים "המטעים"

(יש היפותזה רעה שעבורה ה-loss האמיתי >0, אך ה-empirical loss $=0$)

נבחין ש: $\{S \mid ERM_{\mathcal{H}}(S)\in \mathcal{H}_B\}\subseteq M$, ולכן: $\{S \mid L_{D,f}(ERM_{\mathcal{H}}(S))>\varepsilon\}\subseteq M$

לכן מה שנותר הוא לבחור m המקיים: $D^m(M)<\delta$, וסיימנו!

אז נארגן מחדש את הדגימות ב-M ונכתוב: $M=\bigcup_{h\in\mathcal{H}_B}\{S \mid L_S(h)=0\}$

(ואין בהכרח האיחוד יהיה זר).

נשתמש ב-union bound ונקבל:

$$D^m(\{S \mid L_{D,f}(ERM_{\mathcal{H}}(S))>\varepsilon\}) \underset{\subseteq M \Rightarrow \le D^m(M)\le}{\le} \sum_{h\in\mathcal{H}_B} D^m(\{S \mid L_S(h)=0\})$$

ההסתברות לכל $h\in\mathcal{H}$ לשגות מדגם S עליו h צודקת לגמרי, ∀ לכל x_i. ומתוך ש- $x_i \overset{iid}{\sim} \mathcal{X}$ (בנפרד, באופן בת"ל)

ההסתברות הזו היא ההסתברות ש-h תענה נכון על x_i בנפרד.

נחשב את המשלים: ההסתברות ש-h תטעה על x כלשהם מוגדרת להיות $L_{D,f}(h)$.

אז הסיכוי ש-h תענה נכונה על כל המדגם, בגודל m, הוא: $(1-L_{D,f}(h))^m$, וקיבלנו שלכל h:

$$D^m(\{S \mid L_S(h)=0\})=(1-L_{D,f}(h))^m$$

ובפרט אם $h\in\mathcal{H}_B$ אז $L_{D,f}(h)>\varepsilon$ ונקבל:

$D^m(\{S \mid L_S(h)=0\})<(1-\varepsilon)^m$ לכל $h\in\mathcal{H}_B$. $(\star)$

נסיים את העניין:

$$D^m(\{S \mid L_{D,f}(ERM_{\mathcal{H}}(S))>\varepsilon\}) \le$$
$$\overset{(\star)}{\le} \sum_{h\in\mathcal{H}_B}(1-\varepsilon)^m \underset{\text{(מהנחת הריאליזביליות)}}{<} |\mathcal{H}_B|(1-\varepsilon)^m \le |\mathcal{H}|(1-\varepsilon)^m$$

לבסוף, נשתמש בעובדה ש- $1-\varepsilon \le e^{-\varepsilon}$ ונקבל:

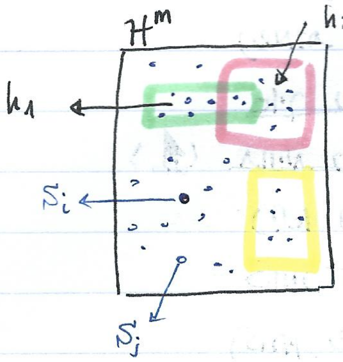
$$D^m(\{S \mid L_{D,f}(ERM_{\mathcal{H}}(S))>\varepsilon\}) < |\mathcal{H}|e^{-\varepsilon m}$$

ו- $|\mathcal{H}|e^{-\varepsilon m}<\delta \Leftrightarrow m\ge \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$,

כלומר עבור $m \ge \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$ נקבל כנדרש. ▨

ה- Union bound

* יעיל מאוד כאשר מתעסקים עם מחלקות היפותזה סופיות.



חיפשנו קב' (בצבעים) שעבורן היפותזות "רעות" שונות מקבלות empirical risk = 0.
(לכל h יש קבוצה אחרת...)
חסמנו את ההסתברות של איחוד הקב' הללו ע"י סכום ההסתברויות שלהן:
מכיוון שההסתברות לקבל קב' כזו בנפרד הוא $< (1-\varepsilon)^m$
אז ההסתברות לטעות הוא להמצא בקב' מהקב' הללו ולכן נרצה להסתכל על ההסתברות של האיחוד שלהן, והוא חסום ע"י סכום ההסתברויות:

$$\sum_{h\in H_B}(1-\varepsilon)^m = |H_B|(1-\varepsilon)^m$$ (שאפשר מסופית H)

אם נחסום את זה מלמעלה ב-δ $\Leftarrow$ נחסום את מה שרצינו.

המשפט היסודי של למידה סטטיסטית:

המשפט היסודי של למידה סטטיסטית:

תהא H מחלקת היפותזות של מסווגים בינאריים עם מימד-$VC = d \leq \infty$.
אזי: H למידת-PAC $\iff d<\infty$. במקרה הזה:

1. ישנם קבועים C_1, C_2 כך שסיבוכיות המדגם של H מקיימת:

$$C_1\frac{d+\log(\frac{1}{\delta})}{\varepsilon} \leq m_H(\varepsilon,\delta) \leq C_2\frac{d\log(\frac{1}{\varepsilon})+\log(\frac{1}{\delta})}{\varepsilon}$$

2. הגבול העליון של סיבוכיות המדגם מושג ע"י לומדן ה-ERM.

לפני שניגש להוכחת המשפט (לא נוכיח אותו לגמרי) ניגע להגדרת ה-framework:

PAC-Framework:

נתקן את הבעיות הבאות (שגורמות לframework לא לבטא נכונה את המציאות):

(1) אנחנו לא ממציאים כאן תוויות רועשות.
(2) הנחת הרליזביליות לא ריאלית. (לא נוכל להגביל את הטבע לבחור רק ממחלקת ההיפותזה.
(3) Misclassification loss - לא ממשק טעויות מסוג I, II. נרצה להגדיר את ה-loss בעצמנו.

Framework

Agnostic PAC :

במסגרת זו נפתרות הבעיות אשר הצגנו מקודם, ובפרט
נשמרת נכונות המשפט היסודי!

<u>איך נגיע אליו?</u>

(1) <u>הגדרות רשמיות</u>: במסגרת זו D תהיה התפלגות מעל $X\times Y$
(ולא מעל X ו-y נקבע עבור x מסוים ע"י f).
כלומר $S \sim D$ מדגם, והאקראיות מתקיימת גם ב-X וגם ב-y.
לפיכך, זוהי התפלגות משותפת.

אנחנו יכולים לתאר את D ב-2 אופנים: כאשר חושבים על D כהתפלגות על $X\times Y$, נסמן (X,Y) משתנים אקראיים שמתפלגים באופן משותף לפי D.

① <u>לפי רגרסיה לוגיסטית</u>: $P(X=x, Y=y) = P(X=x)P(Y=y|X=x)$

כלומר בוחרים את X לפי ההתפלגות השולית, ואז בוחרים את התווית המתאימה y, לפי ההתפלגות המותנית (כמו מקודם...)
אנחנו בנושא קלסיפיקציה בינארית ולכן ההתפלגות השולית Y היא <u>משתנה ברנולי</u>, כלומר קיימת פונקציה: $p: X \to [0,1]$
המקיימת: $P(Y=+1|X=x) = p(x)$. (כלומר ההסתברות תלויה ב-X)
נבחין שאם <u>$p(x)\in\{0,1\}$</u>: אז אנחנו במקרה הדטרמיניסטי. → (כמו x קובעת באופן דטרמיניסטי את התווית שלו)
. <u>אחרת</u>: נדגום את x בדגימות שונות, ונקבל עליו ליבלים שונים.

② <u>לפי מסווג LDA</u>: $P(X=x|Y=y) = P(Y=y)P(X=x|Y=y)$

$P(Y=y)$ הוא הסיכוי לקבל 1 או 0 (כמו להטיל מטבע) כלומר Y <u>משתנה ברנולי</u>. אחרי שבחרנו את הנקודה תג׳ כ: + או -
נבחר את הדגימה מתוך אחת מ-2 התפלגויות (בהתאם ל-Y):
$P(X=x)Y=-1)$ או $P(X=x|Y=1)$.

⇐ מכיוון שעכשיו D התפלגות משותפת מעל $X\times Y$, נגדיר מחדש את ה-loss: Misclassification (general) loss עבור היפותזה $h: X\to Y$:

$$L_D(h) := P_{(x,y)\sim D}\{h(x)\neq y\} \equiv D\{(x,y)|\ h(x)\neq y\}$$

★ כבר אין f. כלומר אין צרך לפונק' האמיתית שהטבע בחר.
כאן, הדבר הקרוב ביותר ל"אמת" שיש לנו הוא ההתפלגות המותנית:
$P(Y=y|X=x)$.

(2) <u>ריאליזביליות</u>: תחת הנחת הריאליזביליות יכולנו להגיע לשגיאה $=0$.
כעת, בשל הרעש, ייתכן שלעולם לא נגיע לשגיאה כזו. לכן נשנה את
מונח ה-accuracy: נרצה שאלג' יחזיר לנו כל החלטה שיהיה במרחק
של לכל היותר ε מהשגיאה האפשרית המינימלית: $\min_{h\in H} L_D(h)$.

• <u>האם המינימום הזה קיים?</u> בתרגול ראינו את Bayes Optimal predictor:
עבור התפלגות D על $X\times Y$:

$$f_D(x) = \begin{cases} 1 & : P(y=1|x) \geq \frac{1}{2} \\ 0 & : \text{אחרת} \end{cases}$$

כלומר זוהי הפונקציה $f_D(x): X\to Y$ שתלויה ב-D (שאינה ידועה).
כלומר היא בעצם Oracle Quantity, ואנחנו לא יכולים
למצוא בעזרתה בפועל (אלא רק להשתמש בו לבדוק אומדן פתרונות מעשיים).
★ תכונה: מסווג זה <u>משיג</u> את ה-loss המינימלי.

ומכך נובע שלכל מחלקת היפותזות $H\subset Y^X$: $L_D(f_D) = \min_{h\in Y^X} L_D(h) \leq \min_{h\in H} L_D(h)$.
ולכן בהחלט נוכל לכתוב $\min_{h\in H} L_D(h)$.

נרחיב את הגדרת למדן Approximately Correct:
יהא $\varepsilon>0$. כל החלטה $h\in H$ הוא AC עם accuracy ε,
(ביחס ל-D) אם: $L_D(h) \leq \min_{h'\in H} L_D(h') + \varepsilon$.

(3) <u>loss כללי</u>:
general loss func: הוא $\ell: H\times Z \to [0,\infty)$ עבור $Z = X\times Y$.
נכתוב ל-$z=(x,y)$, כך: $\ell(h(x),y)$. (שיקולי נוחות). <u>כמה פונק' loss</u>:
(*) <u>Misclassification loss (0-1 loss)</u>:

$$\ell_{0,1}(h,(x,y)) := \begin{cases} 1 & : h(x)\neq y \\ 0 & : h(x)=y \end{cases}$$

(*) <u>$L_D(h)$</u>: (שהגדרנו בעמ' קודם) נבחין שהגדרה שקולה תהיה: →

$$L_D(h) = \mathbb{E}_D[\ell_{0,1}(h,(x,y))]$$

פורמלית
נגדיר מהי פונק' loss כללי: generalization loss (פונק' loss כללית שמושרית ע')

עבור התפלגות D על X×Y, ופונק' loss כללית: $\ell: \mathcal{H}\times Z \to [0,\infty)$
נגדיר את ה-loss של ההיפותזה $h: X \to Y$ ביחס ל-D ול-ℓ:
$L_D(h) := \mathbb{E}_{z\sim D}[\ell(h,z)]$ כאשר $z=(x,y)\in Z$.

ה-accuracy (ביחס ל-loss כללי): יהא $\varepsilon>0$, $h\in\mathcal{H}$ הוא AC
אם ε accuracy ביחס ל-D ולפונק' $\ell: \mathcal{H}\times X\times Y \to [0,\infty)$
ומחלקת ההיפותזות $\mathcal{H}$ אם: $L_D(h) \le \min_{h'\in\mathcal{H}} L_D(h') + \varepsilon$

⇐ הגדרנו את ה-framework, נעבור להגדיר מהי:
Agnostic-PAC Learnability

למידות אגנוסטיק-פאק: מחלקת היפותזות $\mathcal{H}$ היא למידה Agnostic-PAC ביחס לפונק' ה-loss: $\ell: \mathcal{H}\times(X\times Y)\to[0,\infty)$
אם קיימת פונק' $\tilde{m}_{\mathcal{H}}:(0,1)^2\to\mathbb{N}$ ואלגורי' למידה $A:(X\times Y)^m\to\mathcal{H}$ שמקיימת את התכונה:
לכל $\varepsilon,\delta\in(0,1)$ לכל התפלגות D מעל X×Y ולכל $m\ge\tilde{m}_{\mathcal{H}}(\varepsilon,\delta)$:
$$D^m\{S_m \mid L_D(h_S)\le \min_{h\in\mathcal{H}} L_D(h)+\varepsilon\}\ge 1-\delta$$
עבור $S_m=\{(x_i,y_i)\}_{i\in[m]}$ שנדגם iid מ-D ו-$A(S)=h_S$.

תרגיל: Agnostic-PAC Learnability $\Rightarrow$ PAC-Learnability.

משחק הלמידה לאגנוסטיק-PAC:
נקבע accuracy ε ו-confidence δ. ומקבעים $\mathcal{H}\subseteq Y^X$
מחלקת היפותזות ופונק' loss ℓ. משחקים נגד הטבע על random payoff.
1) אנחנו: בוחרים גודל מדגם m, ואלגו' $A:(X,Y)^m\to\mathcal{H}$ ("תלויים" ב-ε,δ)
2) הטבע: בוחר התפלגות D מעל X×Y. (יתכן שתלויה ב-$A,m,\varepsilon,\delta,\mathcal{H}$)
3) נדגום $S=\{(x_i,y_i)\}_{i\in[m]}$ iid מ-D.
4) מקבלים את הפרדיקציה: $A(S)=h_S\in\mathcal{H}$.
5) התשלום הוא לטבע: $L_D(h_S)=\mathbb{E}_{(x,y)\sim D}\,\ell(h,(x,y))$ (מינימיזציה).

המטרה: לייצר h_S PAC (תחת הגדרת ה- general loss) ביחס ל- ε,δ
באופן שלא תלוי בבחירת D של הטבע.

המשפט היסודי של למידה סטטיסטית (עם Agnostic-PAC):

(*) ראשית, נזכיר במשפט את עקרון ה- ERM, ולכן נגדיר אותו (באופן מוכלל):

ERM (general loss): תהא $h: X \to Y$, הסיכון האמפירי של h ביחס לפונק' ה- loss ℓ ולמדגם $S = \{z_i\}_{i=1}^m$ עבור $z_i=(x_i,y_i)$ הוא: $L_S(h) := \frac{1}{m}\sum_{i=1}^m \ell(h,z)$.

לאמן ERM במסגרת Agnostic-PAC: נגדיר את האמן כך:

$$A_{ERM} : S \mapsto \underset{h\in H}{\operatorname{argmin}}\ L_S(h)$$

(*) נדבר עם על חוק המספרים הגדולים:

החוק החלש של המס' הגדולים: יהיו משתנים מקריים ממשיים X_i שנדגמו iid, ויהי $\mu = E(X_i)$. אזי: $\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^m X_i = \mu$

כאשר ההתכנסות היא בהסתברות, ל- $E(X_i)$:

לכל $\varepsilon>0$: $\lim_{m\to\infty} P\left\{\left|\frac{1}{m}\sum X_i - \mu\right| > \varepsilon\right\} = 0$

או באופן שקול, לכל $\delta>0$ קיים $m_0 \in \mathbb{N}$ כך שלכל $m>m_0$:
$P\left\{\left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| > \varepsilon\right\} < \delta$.

ומכאן: $L_S(h) \to L_D(h)$: נבחין שלכל h: $L_S(h) = \frac{1}{m}\sum_{i=1}^m \ell$, אזי מתקיים: $E_D L_S(h) = L_D(h)$, ומחוק המס' הגדולים, עבור S שנדגם iid מגודל m: $\lim_{m\to\infty} L_S(h) = L_D(h)$, בהסתברות.
ולכן לכל $\varepsilon,\delta>0$ קיים $m_0\in\mathbb{N}$ כך שלכל $m>m_0$:
$P\{|L_S(h)-L_D(h)|>\varepsilon\}<\delta$.

פלט ה- ERM

⇐ האם זה אומר ש- $\underset{h\in H}{\operatorname{argmin}}\ L_S(h)$ קרוב ל- $\underset{h\in H}{\operatorname{argmin}}\ L_D(h)$?

למשפט: תהא H מחלקת היפותזות עם מימד-VC $\infty > d$.
H היא למידה Agnostic-PAC $\iff d<\infty$. במקרה זה -

(1) קיימים קבועים C_1, C_2 שסיבוכיות המדגם של H מקיים:

$$C_1 \frac{d+\log(1/\delta)}{\varepsilon} \le m_H(\varepsilon,\delta) \le C_2 \frac{d\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon}$$

(2) העקרון הלמידה מושג ע"י האמן ה- ERM.

ההוכחה
המלאה בספר.

כעת נוכיח את המשפט לא באופן מלא. המטרה היא להראות הוכחה בתאוריה של למידה חישובית, ולקבל אינטואיציה בקשר למימד-VC, שכן החלק שבו נתמקד הוא הטיעון שמחלקה למידת PAC אמ"מ $VCdim(\mathcal{H}) < \infty$, ולא נדבר על החסמים לסיבוכיות המדגם).

הוכחה למשפט היסודי:

למשפט יש 2 כיוונים: $\mathcal{H}$ למידת Agnostic-PAC $\Leftrightarrow VCdim(\mathcal{H}) < \infty$

⇐: בכיוון הנ"ל המשפט בעצם אומר שאם $\mathcal{H}$ היא מחלקת היפותזות עם מימד-VC $= \infty$, אז בוודאי-אפשרי למצוא לה לומד A שהוא Agnostic-PAC (עם confidence δ ו- accuracy ε) זה לא ניתן להשגה ע"י אף אלגוריתם, ולא ע"י דגימות.

בהוכחה של הכיוון הזה כבר דנו בהרצאה קודמת. עיקרי הדברים:

- הוכחנו שאם קיימת $C \subset X$ שמנופצת ע"י $\mathcal{H}$, אז לא קיים ל-$\mathcal{H}$ לומד-PAC שמשתמש ב- $< \frac{|C|}{2}$ דגימות.
- בתרגיל בהרצאה הזו נתבקשנו להוכיח ש: למידות Agnostic-PAC ⇐ למידות PAC. לכן, לא קיים אלגו' Agnostic-PAC ל- $C \subset X$ שמנופצת ע"י $\mathcal{H}$, שמשתמש ב- $< \frac{|C|}{2}$ דגימות.
- מההנחה ש- $VCdim(\mathcal{H}) = \infty$, אז קיימת תת-קב' ב-$X$ שמנופצת ע"י $\mathcal{H}$, ומכל גודל, כלומר לכל קב' C כזו, לא ניתן ללמוד את $\mathcal{H}$ ע"י $\underbrace{m < \frac{|C|}{2}}_{\text{דגימות.}}$ ⇐ לכל m, $\mathcal{H}$ לא ניתנת ללמידה.

⇒: בכיוון הנ"ל המשפט אומר שאם ל-$\mathcal{H}$ מחלקת היפותזות יש מימד-VC$<\infty$ אז היא ניתנת ללמידה, ואף מצא איך: ע"י הלומד האוניברסלי ERM. בנוסף הוא אומר שמה הדגימות שלומד ה-ERM $\underbrace{\text{צריך}}_{ERM_{\mathcal{H}}(S)}$ הוא כמעט אופטימלי.
נקח $h \in ERM_{\mathcal{H}}(S)$ שממזער את השגיאה האמפירית $L_S(h)$ עבור מדגם S.

<u>רוצים להוכיח</u>: $h_S \in ERM_{\mathcal{H}}(S)$ "עובד היטב", כלומר:

$$D^m\{S \in (X \times Y)^m \mid |L_D(h_S) - L_S(h_S)| < \varepsilon\} > 1 - \delta$$

⊛ כמובן שזה לא קורה לכל S, נרצה למצוא S מיוחד עבורה <u>השגיאה האמפירית</u> ($L_S(h)$) קרובה <u>לשגיאה המוחלטת</u> ($L_D(h)$). זה יהיה הקשר שלנו בין מה ש-$ERM_{\mathcal{H}}$ מייצר ($\min_{h \in \mathcal{H}} L_S(h)$) ללמידות-PAC שמייצגת

ע"פ הביטוים במרחק $\ge \varepsilon$ מהביטוי $\min_{h\in H} L_D(h)$. ⇐ נרצה להוכיח התכנסות במ"ש:

תהא סדרת פונק' $f_n: X\to\mathbb{R}$. היא מתכנסת במ"ש ל- $f: X\to\mathbb{R}$ אם

$$\forall\varepsilon>0\ \exists m_0\in\mathbb{N}\ \text{s.t}\ \forall n>m_0,\ \forall x\in X: |f_n(x)-f(x)|<\varepsilon$$

סדרות של מ"מים

⊛ ההוכחה ש- $\min_{h\in H} L_D(h) \sim \min_{h\in H} L_S(h)$ לא קלה:

כמו שראינו: $E[L_S(h)] = L_D(h)$, ולכן מחוק המספרים הגדולים יש התכנסות של $L_D(h) \longrightarrow L_S(h)$. כלומר: לכל D,δ,ε קיים $m_0\in\mathbb{N}$ כך שלכל $m>m_0$: $P\{|L_S(h)-L_D(h)|<\varepsilon\} > 1-\delta$

אבל: m_0 תלוי ב- D וב- h. אנחנו רוצים m_0 שיתפקד אחיד הן בהתפלגויות D והן בהיפותזות $h\in H$. רק תכונה זו תבטיח לנו שכאשר אנו ממזערים את $L_S(h)$ על פני כל $h\in H$, נהיה קרובים למה שממזער את $L_D(h)$ על פני כל $h\in H$. צריכים שכל זה יהיה במ"ש לכל D מכיוון שה' הביטוי שניתן לא אמור להיות תלוי ב- D (למעשה ב"משחקים" אנחנו קובעים m לפני שנקבל לנו D).

אנחנו נחשב על $\{L_S(h)\}_{m=|S|}$ כסדרה של מ"מים, ונרצה שיתקיים: $L_S(h) \xrightarrow[m]{} L_D(h)$ התכנסות במ"ש ביחס ל- $h\in H$ (ה- m שנבחר יבטיח מרחק על כל ההיפותזות) וגם במ"ש ביחס ל- D.

נשתמש במס' כלים:

1. ε-representative training sample (ε-מייצג) : מדגם S נקרא ε-מייצג עבור D,H,l אם לכל $h\in H$: $|L_S(h)-L_D(h)|<\varepsilon$.

⊛ נוח לנו לסווג מדגם כ"טוב".

⇐ תכונה זו מבטיחה ש-h שממזערת את $L_S(h)$ על $h\in H$ (כפי שניתן לנו ע"י ERM) תהיה במרחק $\le\varepsilon$ ממזער של $L_D(h)$ על $h\in H$:

למה: יהי S מדגם $\varepsilon/2$-מייצג ביחס ל- D,H,l. תהא $h_S\in ERM_H(S)$

אז: $L_D(h_S) \le \min_{h\in H} L_D(h) + \varepsilon$.

הוכחה: כתרגיל.

⇐ כלומר ינבע המשפט כפי שרצינו.

2. תכונת ההתכנסות במ"ש (uc) של H: למחלקת היפותזות H יש את תכונת ה-uc אם קיימת פונק' $m_H^{uc}:(0,1)^2\to\mathbb{N}$ כך שלכל $\varepsilon,\delta\in(0,1)$ ולכל התפלגות D על $X\times Y$ מתקיים: $D^m(\{S\in(X\times Y)^m \mid S \text{ מדגם } \varepsilon\text{-מייצג}\}) \ge 1-\delta$

(כלומר במ"ש על D)

תרגיל: אם ל-H יש את תכונת UC עם הפונק' $m_H^{UC}:(0,1)^2\to\mathbb{N}$

אזי H למידה PAC - Agnostic (ע"י ERM) עם סיבוכיות מדגם

$m_H(\varepsilon,\delta) \le m_H^{UC}(\frac{\varepsilon}{2},\delta)$.

⊛ הוכחה בעזרת הלמה + הגדרת תכונת UC.

להוכחת ⇒: נוכיח שאם $vcdim(H)<\infty$ אז H מקיימת תכונת UC.

כלומר, ל-m גדול מספיק (שאינו תלוי ב-D), מדגם שנדגם iid מגודל m

הוא ε-מייצג בהסתברות $\ge 1-\delta$, לכל D.

נגדיר: $F_m^D:(X\times Y)^m\to\mathbb{R}$ ע"י $F_m^D(S):=\sup_{h\in H}|L_D(h)-L_S(h)|$

(*) במילים, פונק' זו ממפה מדגם מגודל m ל"בלבול" הגרוע ביותר שייתכן -

המרחק הגדול ביותר בין $L_S(h)$ ל-$L_D(h)$, על פני כל $h\in H$.

⇐ נרצה להראות שאנו שולטים ב"בלבול" הזה.

(*) נשים לב שהפרמטר S נדגם בצורה אקראית ולכן הוא משתנה אקראי שהתפלגותו

תלויה ב-D וב-m.

נרצה להראות ש-F_m^D קטן: לכל $0<\varepsilon,\delta<1$ קיים $m_H^{UC}(\varepsilon,\delta)\in\mathbb{N}$

כך שלכל התפלגות D מתקיים: $D^m\{F_m^D(S)>\varepsilon\}<\delta$.

⇐ ואז סיימנו ☺: מתקיימת תכונת UC.

נחלק למקרים:

1. → H סופית:

נרצה ללמוד איך להוכיח את המקרה הכללי דרך המקרה הפרטי הפשוט יותר.

הטענה: יהיו δ,ε. אז קיים m_0 כך שלכל $m>m_0$, ולכל D:

$D^m\{F_m^D(S)>\varepsilon\}\le\delta$.

הוכחה:

נבחין: $D^m\{F_m^D(S)>\varepsilon\}=D^m\{S\mid \exists h\in H \text{ s.t. } |L_S(h)-L_D(h)|>\varepsilon\}$

מחסם האיחוד ביטוי זה $\le \sum_{h\in H} D^m\{S\mid |L_S(h)-L_D(h)|>\varepsilon\}$

ומסופיות H, הביטוי האחרון $\le |H|\cdot\max_{h\in H}D^m\{S\mid|L_S(h)-L_D(h)|>\varepsilon\}$.

• מחוק המספרים הגדולים $L_S(h)$ הוא ממוצע של מ"מ שתוחלתם $L_D(h)$,

אז לכל $\varepsilon>0$ שנבחר מתקיים: $D^m\{|L_S(h)-L_D(h)|>\varepsilon\}\xrightarrow[m\to\infty]{}0$.

אנחנו רוצים לחסום את $P(D^m\{|L_S(h)-L_D(h)|>\varepsilon\})$ באופן בל"ת ב-D, h,

כלומר במ"ש, לכן לא נקבל מה שרצינו מחוק המ' הגדולים.

אנחנו רוצים concentration of measure (ריכוז מידע): לחסום את המרחק בין התוחלת לבין הממוצע, ומהי מהירות ההתכנסות.

נשתמש בחסם הופדינג (שהוא מספר אחד החסמים הקיימים לריכוז מידע):

[תהיו $a<\theta_1,\dots,\theta_m<b$ סדרה של מ"מ ב"ת ושווי התפלגות עם $E(\theta_i)=\mu$.
אז לכל $\varepsilon>0$: $P\{|\frac{1}{m}\sum\theta_i-\mu|>\varepsilon\}\le 2e^{(-2m\varepsilon^2)/(b-a)^2}$]

כאן, נגדיר: $L_S(h)=\frac{1}{m}\sum_{i=1}^{m}\theta_i$, $L_D(h)=E(\theta_i)$ (ביחס ל-D!), $\theta_i:=\ell(h,(x_i,y_i))$

ומהופדינג: $D^m\{S \mid |L_S(h)-L_D(h)|>\varepsilon\}\le 2e^{-2m\varepsilon^2}$ (הנחנו ש- $0\le\theta_i\le 1$).

אז: $|H|\max_{h\in H} D^m\{S \mid |L_S(h)-L_D(h)|>\varepsilon\}\le 2|H|e^{-2m\varepsilon^2}$.

וקיבלנו: $D^m\{S \mid \exists h\in H: |L_S(h)-L_D(h)|>\varepsilon\}\le 2|H|e^{-2m\varepsilon^2}$.

ועבור $m\ge\frac{\log(2|H|/\delta)}{2\varepsilon^2}$, נקבל ש: $D^m\{F_m^D(S)>\varepsilon\}<\delta$ כנדרש. ▨

2. <u>למקרה הכללי</u>:

ההוכחה הקודמת לא תעבוד כי $H=\infty$ ולכן אין "$|H|$".

עבור $C\subset X$ סופית נסמן את הצמצום של H ל-C ע"י H_C.

נשתמש במשפט ש: אם $\mathrm{VCdim}(H)<\infty$ אז H מקיימת תכונת-UC.

(וכל כמו שהוכחנו עבור המקרה הסופי ← H למידת Agnostic-PAC כנדרש).

השאלה הזו נעוצה בכמות שבה יכולה H_C לגדול עם $|C|$:

- <u>אם $|C|<\mathrm{VCdim}(H)$</u>: אז יתכן שהיא הקב' המנותצת, כלומר מתקיים: $|H_C|=2^{|C|}$.
- <u>אך אם $|C|>\mathrm{VCdim}(H)$</u>: מהגדרה: $|H_C|\ne 2^{|C|}$. (כלומר $|H_C|<2^{|C|}$).

⇐ אז <u>מה הגודל המקס' של $|H_C|$</u>?

<u>נגדיר</u>: $\mathbb{N}\to\mathbb{N}$: $\tau_H(m):=\max\{|H_C| \mid C\subset X, |C|=m\}$

[מנצלת כמה מהר H_C יכולה לגדול עם $|C|$]

(*) זו הגדרה קומבינטורית והיא תכונה של H: היא' הגודל ביותר של פונק' שאפשר לקבל ע"י הגבלה של H לתת-קב' כשהיא מגודל m.

(**) אינטואיטיבית: עבור H מורכבת יותר נקיים ש-$\tau_H(m)$ תהיה גדולה יותר או תגדל בקצב מהיר יותר.

(*) וראינו שאם $\mathrm{VCdim}(H)=\infty$ אז $\tau_H(m)=2^m$ (כלומר H_C יכולה לגדול אקספ' ב-$|C|$)

גרסה $\mathcal{H}\subset Y^X$. נניח שקיים $m_0\in\mathbb{N}$, $b>0$, $\beta>0$ כך שלכל $m>m_0$:
$\tau_{\mathcal{H}}(m)\le b\cdot m^\beta$, אז נאמר ש- $\mathcal{H}$ גדלה פול׳ ב-m.

נמשיך את ההוכחה ב-2 שלבים:

1. אם $\mathcal{H}$ גדלה פול׳ ב-m, אז $\mathcal{H}$ בעלת תכונת-UC (ולכן למידה PAC ע״י ERM)
2. אם $VCdim(\mathcal{H})<\infty$, אז $\mathcal{H}$ גדלה פול׳ ב-m.

⇐ וסיימנו ☺.

יאללה, נכתוב:

1. צ״ל: $D^m\{F_m^D(S)>m\}<\delta$, באופן בת״ל ב-D.

נשתמש באי״ש מרקוב (ממ״ש אי״ש - $F_m^D(S)\ge 0$ מאקראיות S):

[אם X מ״מ אי״ש אז: $\mathbb{P}\{X>\alpha\}\le E[X]/\alpha$]

מאי״ש מרקוב נקבל שמספיק למצוא סדרת מספרים $\{\alpha_m\}$ עבורה:

$\boxed{E_{D^m}[F_m^D(S)]\le\alpha_m}$ (*) - נשים לב ש-$\{\alpha_m\}$ נסמכת ל-h אך לא ל-D!

ועלינו למצוא חסם אחיד (במ״ט) ל-D, לתוחלת של $F_m^D(S)$.

אם אכן נמצא כזו סדרה $\{\alpha_m\}$ אז: $D^m\{\sup|L_D(h)-L_S(h)|>\epsilon\}=$

$$=\mathbb{P}_{D^m}\{F_m^D(S)>\epsilon\}\underset{\text{מרקוב}}{\le}\frac{E_{D^m}[F_m(S)]}{\epsilon}\underset{(*)}{\le}\frac{\alpha_m}{\epsilon}$$

כלומר בהסתברות $\ge 1-\frac{\alpha_m}{\epsilon}$ מקבלים אימון מגודל m הוא ε-מייצג
אז יש תכונת-UC.

אם נצליח למצוא $\{\alpha_m\}$ כך ששואפת ל-0, אז נקבל שלכל δ אפשר
למצוא m_0 כך שלכל $m>m_0$: $\frac{\alpha_m}{\epsilon}<\delta$. ⇐ ל-$\mathcal{H}$ יש את תכונת-UC!

נמצא $\{\alpha_m\}$:

לא נוכיח → טענה: נגדיר $\sup_{h\in\mathcal{H}}|L_D(h)-L_S(h)|=F_m^D$. אז לכל התפלגות D:

$$E_{D^m}[F_m^D(S)]\le O\left(\sqrt{\frac{\log(\tau_{\mathcal{H}}(2m))}{2m}}\right)+O(m).$$

מההנחה, $\mathcal{H}$ גדלה פול׳ ב-m ולכן קיים m_0 כך שלכל $m>m_0$:

$\tau_{\mathcal{H}}(m)\le b\cdot m^\beta$ עבור $\beta>0$ כלשהו, אז:

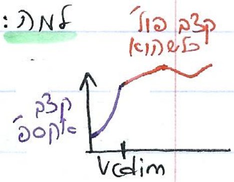
$$E_{D^m}[F_m^D(S)]\le O\left(\sqrt{\frac{\beta\cdot\log(2m)}{2m}}\right)+O(m)\xrightarrow{m\to\infty}0$$

⇐ $\{\alpha_m\}_m=\{\log(m^\beta)\}_m$ מקיים את 6 מה שרצינו - וסיימנו!

(big O אז נפטרנו מהקבוע $\log(b)$)

2. מהגדרה, אם $m\le VCdim(\mathcal{H})$, אז קיימת $C\subset\mathcal{X}$ ש-$\mathcal{H}$ מנתצת.
כלומר לכל m כזה: $\tau_{\mathcal{H}}(m)=2^m$. (זה לא טוב...)
נוכיח בשיעורי הבית את הלמה הבאה:

למה: אם $m\le VCdim(\mathcal{H})$ אז $\tau_{\mathcal{H}}$ אכן גדלה אקס' ב-m.
אך עבור $m>VCdim(\mathcal{H})$ $\tau_{\mathcal{H}}$ גדלה פול' ב-m,
ובפרט במקרה זה $\tau_{\mathcal{H}}(m)\le\left(\frac{em}{d}\right)^d$. (כלומר $\approx$ קבוע $\times\, m^d$).



חוסמת את התוחלת של החריגה הגדולה ביותר האפשרית בין השגיאה הכללית והשגיאה האמפירית (כלו': $\sup_{h\in\mathcal{H}}|L_D(h)-L_S(h)|$)
על S רנדומית iid מ-D.
החסם משתמש ב- $\tau_{\mathcal{H}}(m)$: קצב הגידול של $\mathcal{H}_C$ על $|C|$
(ל- $C\subset\mathcal{X}$).