715|80- הרצאה #4

בעית קלסיפיקציה, נזכיר:

יש domain set X (לרוב $X=\mathbb{R}^d$, היום לא נצטרך להניח זאת), label set Y (אנו עוסקים בבעית קלסיפיקציה פשוטות ולכן זה סיווג בינארי, כלומר $Y=\{0,1\}$)
ו- prediction rule $h: X \to Y$.

האלג' הלומד מקבל $S=\{(x_i, y_i)\}_{i=1}^{m}$ training set
ופולט $h: X \to Y$ prediction rule, learner.
כלומר, learner הוא: $A:(X\times Y)^m \longrightarrow (X\to Y)$ מפה.
$$S \longmapsto h$$

מטרת הלומד היא לפלוט h שידע להחזיר תשובה נכונה על דוגמאות עתידיות. לשם כך, עלינו למצוא ניסוח מתמטי שיתאר כיצד דוגמאות עתידיות מיוצרות.

שאלות עיקריות בלמידה הן:
1. מה ניתן ומה לא ניתן ללמוד?
2. כאשר למידה היא אפשרית, כמה דוגמאות הן צריכה לאימון?
3. כאשר למידה היא אפשרית, כיצד היא נעשית?

⇐ נלמד כאן מהו פורמליזם שמציג תשובות חד-משמעיות לשאלות הללו, בהקשר של batch supervised learning.

Data-generation Model

נניח שיש התפלגות הסתברות D על X.

נניח:

① Simple data generation model: קיימת התפלגות D על X כך ש:
- כל x_i נדגמת לפי D.
- כל x_i נדגמת באופן בת"ל.

⇐ $x_i \overset{iid}{\sim} D$

② נניח שיש פונק' f שהיא true underlying, כלומר, $f(x)=y$

③ $S = (x_i, f(x_i))_{i=1}^{m}$ עבור $x_i \overset{iid}{\sim} D$.

④ כן דוגמא לבדיקה (test sample) $x \in \mathcal{X}$, אף הוא $x \overset{iid}{\sim} D$.

נשווה את <u>המודל הכללי</u>:

* לא הנחנו מאין נלקחות הדוגמאות x_i, ואיזו הסתברות
* הנחנו שקיימת f לינארית: $y_i = f(x_i)$.
* הנחנו שקיים רעש ולכן: $y_i = f(x_i) + z_i$, למשל.

לעומת <u>המודל הזה</u>:

* $x_i \overset{iid}{\sim} D$
* אין רעש: $y = f(x)$. (זה נקרא מודל PAC)
* נוסיף רעש בהמשך

נגדיר את השגיאה של המסווג אותו נלמד:

$$L_{D,f}(h) \overset{def}{=} \underset{x\sim D}{P}[h(x) \neq f(x)] \overset{def}{=} D(\{x \in \mathcal{X} \mid h(x) \neq f(x)\})$$

risk או misclassification error

* נזכור ש-D, f לא ידועות.
* בהמשך נבטל את הבסיס הזה ונטפל גם במקרה מחיר אחרת (שקר, יש כמה סוגים של שגיאות, ונרצה להתייחס לכך)

<u>הבסיס עד כה</u>:

רוצים אלגו' ללמידה supervised batch A שמקבל S (training set), ומחזיר כלל למידה: $h: \mathcal{X} \to \mathcal{Y}$,

<u>כאשר</u>:

קלט:
- $S = (x_i, y_i)_{i=1}^{m}$ ב-$(\mathcal{X} \times \mathcal{Y})^m$
- $\forall i: \ x_i \overset{iid}{\sim} D$
- $\forall i: \ y_i = f(x_i)$ (עבור $f: \mathcal{X} \to \mathcal{Y}$ כלל הלמידה האמיתי שאינו ידוע)

פלט:
- כלל למידה $h: \mathcal{X} \to \mathcal{Y}$, עם שגיאה קטנה: $L_{D,f}(h)$.

☆ אין שליטה על D, f, ו-S רנדומלי. נרצה ש-A יפעל טוב על כל D, f.

נפתח במקרה פשוט זה, וננסה לפתח תאוריה ללמידה. נרצה להבין 2 <u>הגדרות</u>:

1. PAC-learnability ו- sample complexity של מחלקת ההיפותזות $H \subset Y^X$.

2. VC-dimension של מחלקת היפותזות $H \subset Y^X$.

מחלקת היפותזות היא למידת-PAC אם קיימת פונקציה $\tilde{m}_H:(0,1)^2 \to \mathbb{N}$ שמקיימת את התכונה הבאה:
לכל $\varepsilon,\delta \in (0,1)$, לכל התפלגות D על X, ולכל פונק' תיוג $f: X \to \{\pm 1\}$ שעבורה קיימת $h^* \in H$ כך ש- $L_{D,f}(h^*)=0$ - כאשר מריצים את A על יותר מ- $\tilde{m}_H(\varepsilon,\delta)$ דגימות שנדגמות iid מ- D ומתויגות ע"י f, האלגוריתם מחזיר היפותזה $A(S)=h_S$ שמקיימת: בהסתברות של לפחות $1-\delta$ (על S כדגימה מקרית) מתקיים: $L_{D,f}(h_S) \le \varepsilon$.

עבור מחלקת היפותזות H למידת-PAC נגדיר את ה- sample complexity שלה עבור ε,δ נתונים להיות הערך המינימלי של $\tilde{m}_H(\varepsilon,\delta)$ שעבורו ההגדרה הקודמת נכונה. כלומר פונק' ה- sample complexity של H תהיה: $m_H:(0,1)^2 \to \mathbb{N}$

תהא $H \subset \{\pm 1\}^X$ מחלקת היפותזות, ותהא $C \subset X$. נגדיר H_C להיות ההגבלה של H ל- C, כלומר: $H_C = \{h_C \mid h \in H\}$, כאשר $h_C = h|_C : C \to Y$. נגדיר את ה- VC-dimension של H:
$\infty \ge VCdim(H) := \max\{|C| \mid C \subset X \wedge |H_C| = 2^{|C|}\}$.
(כלומר הגודל מירבי של תת-קב' C שעבורה H_C הכי גדולה שאפשר להיות).

למה זה חשוב?

למידת PAC נותנת לנו תורת למידה שלמה שעונה על השאלות המרכזיות בלמידה, אותן הצגנו קודם. The fundamental Theory.

⇐ נקבל: המשפט הבסיסי של למידה סטטיסטית:

⊛ מחלקת היפותזות H היא למידת-PAC $\iff$ $VCdim(H) < \infty$.

⊛ סיבוכיות המדגם של מחלקת היפותזות עם מימד-VC סופי היא בערך: $m_H(\varepsilon,\delta) \sim (VCdim(H) + \log(\frac{1}{\delta}))/\varepsilon$

⊛ כל ה- ERM משיג את המינימום הזה (כלומר אם הלמידה אפשרית, ניתן ללמוד בעזרת כל למידה זה, בערך דיוק קרוב למינימלי).

למידה כמשחק

משחק של 2 שחקנים; אנחנו (מנגנון למידה) מול הטבע,
כאשר התשלום הוא מקרי.
נצטרך להגדיר את המנגנון A בתנאי אי-וודאות (לא יודעים את D,f)
⊛ בתורת החלטות וסטטיסטיקה מקובל להקצות תפקיד קבלת החלטות למשחק.

נקבע את גודל המדגם m, ונתחיל לשחק:
- אנחנו בוחרים אסטרטגיה - $A: (X,Y)^m \to Y^X$
- הטבע יודע את האסטרטגיה שלנו, ובוחר את ההתפלגות D ואת f.
- נדגם מדגם $D \overset{iid}{\sim} S$ בגודל m, והוא מתויג ע"י f.
- המדגם S מועבר ל- A, ואנחנו מחזירים $A(S)=h_S$ כל החלטה.

הגדרת המשחק (1)

התשלום המקרי הוא $L_{D,f}(h_S)$: בתוחלת כמה שגיאות h_S יעשה על מדגם חדש S שנדגם באופן ב"ת ואחיד מ- D, ומקבל תיוג לפי f.
⊛ התשלום אקראי מאקראיות S.
נרצה למצוא A למנגנון שמבטיח loss קטן לכל אסטרטגיה D,f של הטבע.
(כלומר נגיע לאיזה loss לצפות)
האם יש סיכוי לקבל payoff טוב, לכל התנהגות של הטבע?

Probably, Approximately, Correct (PAC)

נשאל 2 שאלות:
1. האם קיימת אסטרטגיה שמשחק שמבטיח הפסד $\varepsilon \in [0,1)$ על ה- loss בהסתברות 1? (כלומר לכל $\overbrace{S}^{L_{D,f}(h_S)}$ חסום ע"י ε) ⟸ לא.
כי תמיד יש הסתברות לקבל מדגם S שלא מייצג את D כמו שצריך ולכן $L_{D,f}(h_S)$ גבוה מאוד, אפילו 1. (אפילו צריקת מטבע <1).

דוגמה:

$X=\{x_1,x_2\}$, ונקבע $\gamma>0$. תהא נק' שלא נצפתה במדגם, אז בה"כ בחרנו לתת + נק' שלא נצפתה לפני.

כעת הטבע משחק ובוחר D המקיימת: $D(\{x_2\})=\gamma$, $D(\{x_1\})=1-\gamma$

ו-f המקיימת: $f(x_1)=f(x_2)=-$ $(\star)$.

נבחין שבהסתברות γ^m, S מכילה רק את x_2. אז מה תהיה $h_S(x_1)$? אין לנו דוגמה, ונבחר $h_S(x_1)=+$.

אך $(\star)$, ולכן טעינו בהסתברות $1-\gamma$, כלומר: $L_{D,f}(h)=1-\gamma$

אם γ קטן מאוד אז ה-loss גבוה מאוד, ובפרט ניתן לבחור γ כך ש: $\varepsilon>1-\gamma$. (למרות ש-γ^m זעירה, ההסתברות קיימת)

למה נכשלנו כרגע? אם γ זעירה, אז x_2 נדירה לעומת x_1 ואנו קיבלנו ב-S רק נק' נדירות, ולא הצלחנו להכליל את D בצורה טובה. מצב כזה תמיד יכול לקרות ולכן התשובה לשאלה 1 - לא.

נרשום קצת את הדרישה: כאשר למידה A יכול לתת לנו כלל החלטה עם loss כשטעה בהסתברות שהיא לכל היותר δ (ל-δ קטנה), אומרים ש-A הוא probably correct learner. ל-δ נקרא confidence.

⇐ כלומר אנחנו יכולים לקבל חסם קבוע (עליון) על ה-loss בהסתברות של $1-\delta$, לא משנה איך הטבע ישחק.

★ האם קיימת אסטרטגיה A שבהסתברות לפחות $(1-\delta)$ תחזיר loss=0 לכל התנהגות של הטבע? ⇐ לא. loss=0 אומר ש: $D\{h_S(x)=f(x)\}=1$, כלומר ש-h_S לא טועה לעולם. אך אם הטבע יספק D שנותן בהסתברות גבוהה מסה זעירה ל-$x\in X$ כשהוא, אז $x\notin S$ בהסתברות גבוהה, כלומר h_S לא יודע איך לסווג את x, והוא לא יכול להכליל את S בצורה מושלמת.

דוגמה:

נקבע $\delta>0$ ו-γ וננסה לחזות בעזרת A נק' $x\notin S$.

בה"כ נניח שבוחרים לחזות בה +. כמקודם $D(\{x_2\})=\gamma$, $D(\{x_1\})=1-\gamma$

בחירת הטבע, ו-$f(x_1)=f(x_2)=-$. אז בהסתברות $(1-\gamma)^m$: $x_2\notin S$.

ל-γ קטן זה גדולה!

$h_S(x_2)=+$ כאמור. אז h_S יטעה בהסתברות γ על $\mathcal{Z}$.

ומכיוון שישרנו את δ, ה-confidence, נקבע את γ לקיים δ לכך $(1-\gamma)^m$,

ונכשלנו. למה נכשלנו ככה? בהסתברות $(1-\gamma)^m$ (גדולה)

מקבלים מדגם "טיפוסי" אך ספציפי נק' נדירה, ולכן לא ניתן

להכליל באופן מושלם.

∎

טענה: לכל confidence δ, לא ניתן למצוא A שמפיק כלל החלטה

h_S עם $L_{D,f}(h_S)=0$ בהסתברות $1-\delta$, לכל D,f.

נשמיט את הכלל מדי ימני: נסתפק בכלל approximately correct

לומר כלל h_S עם $L_{D,f}(h_S)\le\varepsilon$ ל-ε קבוע מראש.

approximately correct learner הוא לומד A עם חסם עליון $\varepsilon>0$

ל-loss שהוא מספק. ל-ε נקרא ה-accuracy.

$\underbrace{\qquad\qquad\qquad}_{\text{על כל התנהגות של הטבע}}$

לומד PAC: לומד A שמפיק כלל רנדומי h_S שעבורו

ה-loss חסום מעל ע"י $\varepsilon>0$ כלשהו, בהסתברות שהיא

לפחות $(1-\delta)$. ללומד מותר לטעות בשגיאה בהסתברות שהיא

לכל היותר δ. נקרא ל-A: probably approximately correct

עם ε accuracy ו-$1-\delta$ confidence.

ובאופן מתמטי: $A: S\to h_S$ לומד-PAC (עם δ,ε) אם

לכל D,f: $\mathbb{P}_{D^m}\{S\in(\mathcal{X}\times\mathcal{Y})^m \mid L_{D,f}(h_S)\le\varepsilon\}>1-\delta$

[כאשר $((x_i,y_i))_{i=1}^m$ ו-x_i iid מ-D]

אז אנחנו יכולים לקוות לבנות לומד PAC, ~~ועכשיו את ההגדרות~~

המשחק: • m לא קבוע; פרמטרי המשחק יהפכו להיות δ,ε.

• אנו נחליט על m באסטרטגיה שלנו, ולכן כאשר נגדיר A,

למעשה נגדיר לרצף A_m (למדנים לכל m):

כאשר $A_m:(\mathcal{X}\times\mathcal{Y})^m\to\mathcal{Y}^{\mathcal{X}}$.

הגדרת המשחק (*)

נקבע $\varepsilon>0$: accuracy , $\delta>0$:confidence ונשחק משחק נגד הטבע עם random payoff.

- בתורנו נבחר את m (sample size), ולומד: $A: (X\times Y)^m \to Y^X$. נניח ש-m, A יכולים להיות תלויים ב-δ, ε.
- הטבע יודע את A,m, ובוחר אסטרטגיה שמורכבת מהתפלגות D על X ו- $f: X\to Y$ פונק' תיוג. כלומר אסטרטגיה שתלויה ב-$\delta, \varepsilon, m, A$.
- המשחק ממשיך: נדגום $\underbrace{m}_{\text{מדגם } S:}$ דגימות iid מ-D, ונתייג באמצעות f
- נחשב את $h_S = A(S)$.
- נאמר את ה-payoff: $L_{D,f}(h_S)$ (רנדומי כי S מקרי ולכן גם h_S).

נשחק את המשחק (עם אותן האסטרטגיות) ∞ פעמים, ונחשב את ההתפלגות על פני המדגמים האפשריים של המ"מ: $\{S\sim D^m \mid L_{D,f}(h_S)\le\varepsilon\}$ (כמה מהמדגמים קיבלנו: $loss \le \varepsilon$).
אם ההסתברות הזו $> 1-\delta$ (כלומר A הוא לומד-PAC) כנגד G אסטרטגיה של הטבע, נאמר שניצחנו (ביחס לפרמטרים δ, ε).

<u>מהו m?</u> נרצה להסתפק ב-m מינימלי ככל האפשר, אך שהיה עדיין מספיק גדול כדי שנוכל למצוא A שהוא לומד-PAC. אכן יש tradeoff. ניתן לנתח אותו ע"י קיבוע 2 פרמטרים ונימצא הערך הטוב ביותר עבור השלישי, תחת ההנחות (פרמטרים: $\delta, \varepsilon, m(\varepsilon,\delta)$)

⇐ אבל, אין אפשרות לנצח את הטבע בבחירה כזו לכל δ, ε...
<u>למה?</u> אם לא יודעים כלל את D,f, לא נחשב כמה גדול יהיה m
לא נוכל למצוא h_S שגבוה confident מספיק ו- accurate מספיק.

כד' להפריך את ההנחה במשחק (1) נזקקנו ל-$|X|=2$. כעת נניח $|X|=\aleph_0$.

<u>דוגמא:</u>
נניח שבתורנו בחרנו m, ועל כל נק' $x\notin S$ ננבא $g(x)$. נזכור שהטבע יודע את הבחירות אלה. והוא בוחר $C\subset X$ כך ש: $|C|>2m$, D התפלגות אחידה על C ו- $f(x)=-g(x)$ לכל $x\in X$.

כי D על C.

נמשך במשחק: • תידגם S. נחשב: $Supp(S) \subseteq C$. מתקיים: $|Supp(S)| \le m$, ומכך ש-D אחידה על C מתקיים - $D(\{x \in X \setminus Supp(S)\}) \ge 1/2$ כלומר ההסתברות ש נק' שאיננה מ-S לפי D היא לפחות $\frac{1}{2}$.

• $h_S = A(S)$

• נחשב את ה-loss: כל ההחלטה יחזה $g(x)$ על נק' שאין ב-S, והן קטנות הסתברות $\le \frac{1}{2}$ להבחר, ויטעה מכיוון שהטבע החליט שעבורן $f(x) = -g(x)$. לכן $L_{D,f}(h_S) \ge \frac{1}{2}$, לכל S training set שיבחר.

⇐ לכל למדן A שנו הטבע בוחר תצורה עבורה ה-$loss \ge \frac{1}{2}$. גם אם נגדיל את m זה לא יעזור - הטבע יבחר קב' C גדולה יותר, והוא יכול כי $|X| = \aleph_0$.

<u>למה נכשלנו ככה?</u> הטבע בחר D ו-f בעולם, ואנו ניסינו להגיע לניקוי ממדגם קטן מדי מכל הפונק' האפשריות. לכן אם מרחב הפונק' האפשריות ל-f גדול מדי - לא נוכל לייצר למדן PAC. ⇐ העקרון של משפט אין ארוחות חינם (="בלי מידע מקדים על הפונק' f אותה אנו מנסים ללמוד - הלמידה בלתי-אפשרית") ▨

<u>משפט אין ארוחות חינם</u>: תהא X קב' הדגימות אינסופית, ונקבע $\varepsilon < \frac{1}{2}$. אזי [illegible] כך שלכל למדן A ו-S training set בגודל m קיימת התפלגות $D(x)$ על X ופונק' תיוג $f: X \to Y$ כך שבהסתברות $> \delta$ על ייצור ה-training sample S בגודל m שנדגם iid מ-D מתקיים $L_{D,f}(h_S) \ge \varepsilon$ ל-$h_S = A(S)$.

כלומר כדי ללמוד אנו חייבים להניח מידע מקדים כלשהו על קב' ההיפותזות $H \subseteq Y^X$ ממנה f נבחרת.

<u>נתבונן ב-2 מקרים:</u>

(1) <u>Realizable</u>: הטבע חייב לבחור $f \in H$. למעשה מספיק שיבחר פונק' f שהיא <u>D-almost surely</u> זהה ל-$h^* \in H$. כלומר: כלומר- $\exists h^* \in H \text{ s.t. } D(\{f(x) = h^*(x)\}) = 1$.

טעם הלומד יחזיר: $A(S) = h_S : (X \times Y)^m \to \mathcal{H}$.

<u>מערכת שאלות חדשות</u>:

1. מהי $\mathcal{H}$ "מספיק קטנה" שעבורה ניתן למצוא לומד - PAC?
מהי $\mathcal{H}$ "גדולה מדי" שעבורה לא ניתן למצוא לומד - PAC?

2. נניח שיש לנו $\mathcal{H}$ "מספיק קטנה". אזי לכל δ, ε יש אסטרטגיה A, m שהיא PAC (עם δ, ε), לא משנה איזו אסטרטגיה הטבע בוחר. כלומר לכל δ, ε יש מס' מינימלי של דגימות אימון! האם אנחנו יכולים לאפיין את הפונק' הזו - $m_{\mathcal{H}}(\varepsilon, \delta)$ - שאומרת את מס' הדגימות המינימלי שעבורו ניתן למצוא לומד - PAC עבור δ, ε? (האם יש קשר בין הסיבוכיות של $\mathcal{H}$ לבין מס' הדגימות שנדרש כדי ללמוד אותה?)

3. נניח שיש לנו $\mathcal{H}$ "מספיק קטנה", האם אפשר לכתוב את הלומד שבאמת מצליח למצוא פונק' מ-$\mathcal{H}$, <u>באופן מפורש</u>? כמה דוגמאות m הוא צריך כדי להיות לומד PAC לכל אסטרטגיה של הטבע? האם אפשר למצוא באופן מפורש את הלומד <u>היעיל ביותר</u> עם מס' דגימות מינימלי ($m_{\mathcal{H}}(\varepsilon, \delta)$)?

עכשיו שוב את <u>הגדרת המשחק</u>:

נקבע את $\varepsilon > 0$ - accuracy, $\delta > 0$ - confidence, ואת $\mathcal{H} \subset Y^X$ מחלקת היפותזות. משחקים משחק נגד הטבע עם random payoff.

הגדרת משחק (3)

- בתורנו נבחר m, ו-$A : (X, Y)^m \to \mathcal{H}$ (יכולים להסתמך על δ, ε).
- הטבע יודע את האסטרטגיה (ואת $\delta, \varepsilon, \mathcal{H}$) שלנו ובוחר התפלגות $\mathcal{D}$ על X ופונק' תיוג $f \in \mathcal{H}$ (או פונק' f שהיא $\mathcal{D}$-almost-surely זהה לפונק' מ-$\mathcal{H}$).
- דוגמים $S \overset{iid}{\sim} \mathcal{D}^m$, ומתייגים אותו בעזרת f.
- $A(S) = h_S$.
- מחשבים את ה-payoff <u>המקרי</u>: $L_{\mathcal{D},f}(h_S)$ (כי S מקרי).

☆ נחשב: $\mathbb{P}_{\mathcal{D}^m}\{S \sim \mathcal{D}^m \mid L_{\mathcal{D},f}(h_S) \le \varepsilon\} = P$ על ניסויים חוזרים (S שונים). אם $P > 1-\delta \Rightarrow$ נאמר <u>שהצלחנו</u>, כלומר A הוא לומד - PAC עם הפרמטרים δ, ε.

ראינו שכאשר $|X|=\infty$ ו- $H=Y^X$ לא ניתן ללמוד. אבל

אם H "קטנה מספיק" אז גם כאשר $|X|=\infty$ ניתן ללמוד-PAC!

דוגמא: $X=\mathbb{R}$, $Y=\{0,1\}$. נגדיר את מחלקת ההפרדות

$H_{th}=\{x\mapsto h_\theta(x) \mid \theta\in\mathbb{R}\cup\pm\infty\}$: Threshold functions

$$h_\theta(x)=\begin{cases} x\le\theta: & 0\\ x>\theta: & 1\end{cases} \qquad (h_{-\infty}(x)\equiv 1,\ h_{\infty}(x)\equiv 0)$$

★ תחת הנחת ה-realizability: לא נקבל קונפיגורציה מהצורה

[נק' שמתויגות 0 / נק' שמתויגות 1]

שבה

הטענה שנראה: לכל δ,ε וללא תלות באסטרטגיה של הטבע (בחירת התפלגות D על $\mathbb{R}$, ו- $f\in H_{th}$) - אנחנו נצליח ללמוד-PAC עם δ,ε רצויים.

נשחק:

• עבור S training samples, נבחר את θ להיות: $\theta_{alg}=\max_{y_i=0}(x_i)$

ונבחר את $h_{\theta_{alg}}\in H_{th}$, ואם כל S מתויג 1, נבחר $\theta_{alg}=\infty$

ואם כל S מתויג 0 נבחר $\theta_{alg}=-\infty$.

משפט - נקבע $\delta,\varepsilon>0$. אם $m\ge\frac{1}{\varepsilon}\log(\frac{1}{\delta})$ אז לכל התפלגות D על $\mathbb{R}$, ולכל בחירה של $f_\theta\in H_{th}$ בהסתברות לפחות $1-\delta$ (על $S\overset{iid}{\sim}D$, וגם $m=|S|$) ה-loss: $L_{D,f}(h_{\theta_{alg}})$ הוא לכל היותר ε.

הוכחה:

• הטבע בוחר D על $\mathbb{R}$ ומציין f_0 $f_\theta\in H_{th}$.

★ נבחין ש- $\theta_{alg}<\theta$. כי אם הטבע מחליט θ אז המדגם ניתן לתייג 0 רק משמאל ל- θ ו- 1 רק מימין ל- θ.

⇐ כלומר אנחנו $\theta_{alg}<\theta$, ולכן נסווג נכון עבור $x\le\theta_{alg}$ ועבור $x>\theta$, ונטעה עבור $\theta_{alg}<x<\theta$.

(נרצה שזה יקרה בהסתברות $\le\delta$ ואז נהיה למדן PAC)

לא הוכחה אמיתית, [illegible] ולא נכנס לזה, אלא נבין את הרעיון.

נחלק למקרים:

(1) $D(\{x\mid -\infty<x<\theta\})<\varepsilon$: אז $D(\{x\mid x\in(\theta_{alg},\theta)\})<\varepsilon$ (כי זאת תת-קבוצה), ולכן

$L_{D,h_\theta}(h_{\theta_{alg}})<\varepsilon$ בהסתברות 1: לא משנה מהי S או m. ☺

(2) $D(\{x\mid -\infty<x<\theta\})\ge\varepsilon$: אז אפשר לבחור θ' כך ש:

$D(\{x\mid \theta'<x<\theta\})=\varepsilon$. אם יש לנו $(x,y)\in S$ כך ש: $x\in[\theta',\theta]$

אז ה-loss $\le\varepsilon$ וההסתברות לא לקבל דגימה כזו היא $(1-\varepsilon)^m$.

(כי ה-x הנ"ל יוביל לנו ש- $\theta_{alg}\in[\theta',\theta]$ ולכן השגיאה שלנו היא בקטע הזה, כלומר $\le\varepsilon$)

ובגלל שמתקיים: $1-\varepsilon \le e^{-\varepsilon}$ נקבל ש: $(1-\varepsilon)^m \le e^{-\varepsilon m} < \delta \Leftrightarrow m \ge \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)$.
כלומר בהסתברות $\ge 1-\delta$ השגיאה אכן $\ge \varepsilon$. $\Leftarrow$ A למידה-PAC.

ראינו שיש מחלקת היפותזות שעבורן יש למידה-PAC (כלומר ניתן ללמוד אותן). מחלקות אלו נקראות **PAC-learnable Hypothesis classes**.

ניתן לבחור גודל מדגם סופי

מתי יש למידות-PAC? נגלה

(1) עבור מחלקת היפותזות סופית: יש למידה פשוטה שמוצאת מנבא מינימום (מנבא) באושר מחלקת ההיפותזות סופית! (עכשיו שאם עבור כמה שאינן סופיות): ננסה להבין כמה שיותר (למזער את ה-loss) על ה-training set. $\Leftarrow$ עקרון ה-ERM:

יהא $S=\{(x_i,y_i)\}_{i=1}^m$ training set, אז ה-empirical risk של פונקציה מועמדת $h\in\mathcal{H}$ הוא: $L_S(h)=\frac{1}{m}\left|\{i \mid h(x_i)\neq y_i\}\right|$

האלגוריתם מחזיר תמיד את h על S שממזער את ה-empirical risk, כלומר: $A_{ERM}: S\mapsto \underset{h\in\mathcal{H}}{\mathrm{argmin}}\{L_S(h)\}$. לאלגוריתם זה נקרא **למידה-ERM** ($ERM_{\mathcal{H}}$).

☆ נבחין ש: $L_S(h)\ge 0$, ומחלקת ההיפותזות סופית, ולכן קיים מינימום argmin.
כמו כן, מהנחת ה-realizability, לכל $f\in\mathcal{H}$ שהפונקציה מחזירה מתקיים: $L_S(f)=0$, ולכן ניתן להשיג $L_S(h)=0$.

☆ כלומר תחת ההנחה (realizability) למידה ה-ERM תמיד יחזיר $h\in\mathcal{H}$ עם $h(x_i)=y_i$ לכל $i\in[m]$. לh כזה נקרא **consistent** כי הוא קונסיסטנטי עם ה-training sample

$\Leftarrow$ נוכיח בהמשך שלכל $\mathcal{H}\subset\mathcal{Y}^X$ סופית אז היא למידה-PAC. באמצעות עקרון ה-ERM, עם סיבוכיות מדגם $\le \frac{1}{\varepsilon}\cdot\log\left(\frac{|\mathcal{H}|}{\delta}\right)$. כלומר:

משפט: יהא X sample space, $\mathcal{Y}=\{0,1\}$, ו- $\mathcal{H}\subset\mathcal{Y}^X$ סופית $|\mathcal{H}|<\infty$. (נוכיח בהרצאה הבאה)
נקבע $1>\varepsilon,\delta>0$. אזי אם $m\ge\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$, אז לכל D,f מתקיים: ש $L_{D,f}(ERM_{\mathcal{H}}(S))\le\varepsilon$, בהסתברות $\ge 1-\delta$ (עבור S,m).

☆ log עולה לאט (גם אם $\mathcal{H}$ ענקית) ולכן זה אלגוריתם אופטימלי ללמידה (ג"ג).

הקשר בין m, δ, ε: • אם קבענו את δ, ε אז: $m_H(\varepsilon,\delta) \le \dfrac{\log\left(\frac{|H|}{\delta}\right)}{\varepsilon}$

כלומר ככל ש-H מורכבת יותר (או ש-ε או δ קטן יותר) - צריך יותר דוגמאות.
והמדד למורכבות הוא $|H|$. רואים שמס' הדוגמאות עולה באופן
פרופורציונלי ל- $\frac{1}{\varepsilon}$, אך באופן פרופורציונלי ל- $\log\left(\frac{1}{\delta}\right)$.

• אם קבענו את δ, m אז: $\varepsilon \le \dfrac{\log\left(\frac{|H|}{\delta}\right)}{m}$

• אם קבענו את ε, m אז: $\delta \le \dfrac{|H|}{e^{m\varepsilon}}$

שאלות הבשלות:

1. האם $m_H(\varepsilon,\delta) \le \frac{\log(|H|/\delta)}{\varepsilon}$ באופן הדוק? (האם קיים למידן PAC ל- ε, δ (כל) שמצליח ללמוד בפחות דוגמאות?)
2. מה קורה כאשר נניח מציאות רעש?
3. מה קורה עבור מחלקת השערות שאינה סופית?

תשובה:

3. ראינו ש- H_{th} (שהיא אינה סופית) היא למידת-PAC, וסיבוכיות המדגם היה: $m_{H_{th}}(\varepsilon,\delta) \le \frac{\log(1/\delta)}{\varepsilon}$, שהושג ע"י שימוש אם משתמשים ב- $ERM_{H_{th}}$.

כלומר למרות שזו מחלקה ∞, היא פשוטה. נרצה דרך אחרת לאמוד את מורכבות של מחלקת השערות (למשל יחס סדר קיים מבחינת מורכבות).
ואז נמצא פה שסופי - המחלקה למידת PAC, ואחרת - לא.

דוגמה: $X=\mathbb{R}$, $H=\{h_{a,b,c} \mid a<b<c\in\mathbb{R}\}$ כאשר:

$$h_{a,b,c}(x) = \begin{cases} 1 & x\in[a,b] \vee x\ge c \\ 0 & \text{אחרת} \end{cases}$$

(כלומר:  ([תווית 0, תווית 1])

אז זו מחלקה "מורכבת יותר" מאשר H_{th}.

הגדרה $C\subset X$ ו- $h: X\to Y$ השערה. נסמן את הצמצום של h על C
כ: $h_C: C\to Y$ כך ש- $h_C(x)=h(x)$ לכל $x\in C$.

נניח ש-H מכילה את כל h_C על $C\in X$ (כלשהי) בגודל m. אז לא נוכל למצוא
למידן-PAC שמשתמש במס' דוגמאות $\le \frac{m}{2}$.

כמו מקודם - לא הוכחה אמיתית.

למה? נניח שקיימת $C\subset\mathcal{X}$ מגודל m כנ״ל. משחקים במשחק נגד הטבע ובחרנו $\frac{m}{2}$ דוגמאות כאלה, ולמדן A. אז הלומדן שלנו חוזה עבור $x\in\mathcal{X}$ שלא נמצא ב-S: $g(x)$ (כלומר מייצר $g:\mathcal{X}\to\mathcal{Y}$ ונדאג רק ש- $h_S\in\mathcal{H}$).

הטבע בוחרת D מעל C אחידה, ובוחרת את f להיות $g(x)$- לכל $x\in C$ (אפשרי כי D נתמכת על C, ולטבע לא אכפת איך f מוגדר על $x\notin C$). זה צעד חוקי למרות הנחת ה-realizability) כי הנחנו שכל $h:C\to\mathcal{Y}$ מתקבלות מהגבלה של $\mathcal{H}$ על C.
כלומר הטבע עדיין לא ~~[illegible]~~ $h^*\in\mathcal{H}$ כך ש- $f=h^*|_C$.
והלומדן שלנו ״יפשל״ כי $\frac{1}{2}$ מ-C הוא לא ראה, ויש הסתברות $\frac{1}{2}$ לקבל x שלא ראה, ולכן הוא יטעה על ה-$\frac{1}{2}$ הזה.
כלומר כל עוד ש-$\mathcal{H}$ מכילה קב׳ C מגודל $2m$ שמקיימת: $(C\to\mathcal{Y})=\{h_C\mid h\in\mathcal{H}\}$ אז לא נוכל ללמוד את $\mathcal{H}$ (למידת PAC) בהינתן training sample מגודל m.

⇐ הגודל המירבי של C שמקיימת $(C\to\mathcal{Y})=\{h_C\mid h\in\mathcal{H}\}$ הוא קריטי:
(i) הוא מספק לנו חסם תחתון על $m_{\mathcal{H}}$ (sample complexity)
(ii) אם הגודל המירבי הוא ∞ (לכל $n\in\mathbb{N}$, $\mathcal{H}$ מכילה קב׳ כנ״ל עם גודל $> n$) אז $\mathcal{H}$ אינה למידת-PAC!

תהא $C=\{x_1,\dots,x_{|C|}\}$ ונגדיר $\mathcal{H}_C$ ההגבלה של $\mathcal{H}$ על C. (כלומר $\mathcal{H}_C=\{h_C\mid h\in\mathcal{H}\}$). נעבוד על $\mathcal{Y}=\{\pm1\}$, כך שנוכל לייצג כל h_C ע״י וקטור $(h(x_1),\dots,h(x_{|C|}))\in\{\pm1\}^{|C|}$. כלומר המס׳ של וקטורים כאלו הוא $2^{|C|}$, אז: $|\mathcal{H}_C|\le 2^{|C|}$.
אם מתקיים ש: $|\mathcal{H}_C|=2^{|C|}$ עבור $C\subset\mathcal{X}$ נגיד ש: $\mathcal{H}$ shatters C.
כלומר $\boxed{\mathcal{H}_C=\{h_C\mid h\in\mathcal{H}\}=\mathcal{Y}^C}$
(ואז מהחסם הקודם לא נוכל ללמוד אותה עם מס׳ דוגמאות $< 2|C|$).

מימד VC של מחלקת היפותזות: $VC_{dim}(\mathcal{H})=\max\{|C|\mid \mathcal{H}\text{ shatters }C\}$

☆ נחשוב על זה כך - הגודל המירבי של קב׳ C ב-$\mathcal{X}$, ש-$\mathcal{H}$ לא נותנת עליה שום מידע קודם לגבי הלייבל (אם ראינו כל זה אמור לעזור לנו להשיג h_S מתאימה אבל עבור נק׳ ב-C, נקבל בממוצע של כל מחלקת ההיפותזות, כלומר הוא לא עזרה לנו בכלל)

כדי לחשב מימד VC - <u>צריך להראות 2 דברים</u>:

1. קיימת קב' C מגודל d שמנופצת ע"י H.
2. כל קב' C מגודל $\geq (d+1)$ לא מנופצת ע"י H.