תרגיל בית 2 – מבוא ללמידה חישובית

מגיש: נתן בלוך

Theory Questions.

שאלה 1. Singletons.

נתון $\mathcal{X}$ תחום דיסקרטי, ונסמן את מחלקת ההיפותזות הנתונה ב- $\mathcal{H}_{Singleton}$, כפי שמוגדרת בשאלה.
הנחת ה-$realizability$ במקרה זה גוררת במקרה זה כי ההיפותזה האמיתית, שנסמנה ב-f, עושה $labeling$ של 0 לכל התחום $\mathcal{X}$, מלבד **אולי** לנקודה אחת בו.
נסמן ב-S את ה-$training\ set$, וכן נסמן –

$$S = \{(x_1, y_1), \dots, (x_n, y_n)\} \quad \boldsymbol{as} \quad \forall i \in \{1, \dots n\}:\ x_i \in \mathcal{X}\ and\ y_i \in \{0,1\}$$

נתאר אלגוריתם שיממש את כלל ה-ERM תחת ההנחה של $realizability$.
האלגוריתם הינו באופן הבא –

1. נעבור על כל הדגימות $(x_i, y_i) \in S$, עבור $i = 1, \dots, n$:
   - אם $y_i = 1$, נחזיר את הפונקצייה $h_{x_i} \in \mathcal{H}_{Singleton}$.
2. אחרת, מתקיים ש- $y_i = 0$ לכל $i = 1, \dots, n$, ובמקרה זה נחזיר את $h^- \in \mathcal{H}_{Singleton}$.

נכונות. תחת הנחת ה-$realizability$ קיימת $f \in \mathcal{H}_{Singleton}$, כך ש – $f(X) = Y$. ראינו בהרצאה כי למעשה תנאי זה גורר דטרמיניסטיות של Y בהנתן X. בנוסף, מכך ש – $f \in \mathcal{H}_{Singleton}$, נסיק כי ישנו **לכל היותר** $x \in \mathcal{X}$ **אחד ויחיד** עבורו $f(x) = 1$, ולכל שאר התחום $\mathcal{X}$, מתקיים ש – $f(x) = 0$.
כמו כן, מהנחת ה – $realizability$, נסיק כי מתקיים – $e_S(f) = 0$. על כן, **על מנת להוכיח כי האלגוריתם שתואר הינו אכן אלגוריתם ERM**, נדרש להראות כי מתקיים – $e_S(A(S)) = 0$.
כאשר A מייצג את האלגוריתם שתואר, וכן כאמור הקלט לאלגוריתם הינו S, אוסף של n דגימות מקריות.
נחלק למקרים לפי השלב בו האלגוריתם הסתיים –

- מקרה 1 – האלגוריתם הסתיים בשלב הראשון(1). לכן, לפי האלגוריתם, קיימת דגימה (x_i, y_i) עבורה $y_i = 1$. מהנחת ה-$realizability$, מתקיים כי – $y_i = f(x_i) = 1$, וכן מכך ש- $f \in \mathcal{H}_{Singleton}$, נסיק כי בהכרח $f = h_{x_i}$. כמו כן, במקרה זה האלגוריתם מחזיר את h_{x_i}, ולכן –
$$e_S(A(S)) = e_S(h_{x_i}) = e_S(f) = 0$$
כלומר השגיאה האמפירית של $ERM(S)$ במקרה הנ"ל הינה 0, כנדרש על מנת להראות שזה אכן אלגוריתם ERM.
- מקרה 2 – האלגוריתם הסתיים בשלב השני(2). לכן, לפי האלגוריתם, לכל הדגימות ב-S, מתקיים ש – $y_i = 0$ לכל i, ומהנחת ה-$realizability$, מתקיים $y_i = f(x_i) = 0$. כמו כן, במקרה זה האלגוריתם מחזיר את h^-, שעבורה לכל $x \in \mathcal{X}$, מתקיים $h^-(x) = 0$, ובפרט נבחין כי h^- ו-f מזדהות על הקבוצה $\{x_1, \dots, x_n\}$, ולכן ניתן למעשה להסיק כי השגיאות האמפיריות זהות, ולכן –
$$e_S(A(S)) = e_S(h^-) = e_S(f) = 0$$
כלומר השגיאה האמפירית של $ERM(S)$ במקרה הנ"ל הינה 0, כנדרש על מנת להראות שזה אכן אלגוריתם ERM.

ובכל מקרה, קיבלנו כי השגיאה האמפירית של ההיפתוזה שהאלגוריתם מחזיר הינה 0, ולכן האלגוריתם שלנו הוא אכן אלגוריתם *ERM*.

שאלה 2. PAC In Expectation.

נוכיח את הטענה הבאה –

$$\mathcal{H} \;\; is \;\; PAC \; learnable \quad \boldsymbol{if \; and \; only \; if} \quad \mathcal{H} \;\; is \;\; PAC \; learnable \; in \; expectation$$

תחת הנחת ה-$\boldsymbol{realizability}$.

פיתרון.

בכיוון $\Leftarrow$, נניח כי $\mathcal{H}$ היא $PAC\ learnable$ בתוחלת. מהגדרה, קיים אלגוריתם לומד A, ופונקצייה $N(a)$, שמוגדרת ע"י $N(a) : (0,1) \to \mathbb{N}$, כך שלכל $a \in (0,1)$, ולכל התפלגות P, בהנתן מדגם S, עבורו $|S| > N(a)$, מתקיים –

$$\mathbb{E}\left[e_P\big(A(S)\big)\right] \le a$$

נרצה להראות כי $\mathcal{H}$ היא $PAC\ learnable$, ולשם כך נדרש להוכיח כי קיים אלגוריתם לומד A^*, ופונק' $N^*(\varepsilon,\delta)$, שמוגדרת ע"י $N^*(\varepsilon,\delta) : (0,1) \times (0,1) \to \mathbb{N}$, כך שלכל $\varepsilon,\delta \in (0,1)$, ולכל התפלגות P, בהנתן מדגם S, עבורו $|S| > N^*(\varepsilon,\delta)$, מתקיים –

$$\mathbb{P}\left[e_P\big(A(S)\big) > \varepsilon\right] \le \delta$$

ואכן, נוכיח כי עבור $A^* := A$, כלומר **אלגוריתם זהה, ועבור** $\boldsymbol{N^*(\varepsilon,\delta)}$ **המוגדרת ע"י –** $\boldsymbol{N^*(\varepsilon,\delta) := N(\varepsilon \cdot \delta)}$, מקיימת את התנאי הדרוש.

יהיו $\varepsilon,\delta \in (0,1)$, תהי התפלגות P, ויהי מדגם S, עבורו $|S| > N^*(\varepsilon,\delta)$. **נוכיח** כי מתקיים –

$$\mathbb{P}\left[e_P\big(A(S)\big) > \varepsilon\right] \le \delta$$

כיוון שהנחנו $|S| > N^*(\varepsilon,\delta) = N(\varepsilon \cdot \delta)$, ניתן להסיק מכך שהמחלקה $\mathcal{H}$ היא $PAC\ learnable$ בתוחלת, עבור $a = \varepsilon \cdot \delta$, נקבל –

$$\mathbb{E}\left[e_P\big(A(S)\big)\right] \le \varepsilon \cdot \delta$$

מכאן, נקבל –

$$\frac{\mathbb{E}\left[e_P\big(A(S)\big)\right]}{\varepsilon} \le \delta$$

ועל ידי שימוש באי-שיוויון מרקוב נקבל –

$$\mathbb{P}\left[e_P\big(A(S)\big) > \varepsilon\right] \le \mathbb{P}\left[e_P\big(A(S)\big) \ge \varepsilon\right] \underbrace{\le}_{Markov} \frac{\mathbb{E}\left[e_P\big(A(S)\big)\right]}{\varepsilon} \le \delta$$

וכאמור קיבלנו –

$$\mathbb{P}\left[e_P\big(A(S)\big) > \varepsilon\right] \; \le \; \delta$$

ובעצם הוכחנו כי המחלקה $\boldsymbol{\mathcal{H}}$ היא $\boldsymbol{PAC\ learnable}$, לפי ההגדרה, כנדרש.

בכיוון $\Rightarrow$, נניח כי $\mathcal{H}$ היא $PAC\ learnable$. מהגדרה, קיים אלגוריתם לומד A, ופונקצייה $N(\varepsilon,\delta)$, שמוגדרת ע"י $N(\varepsilon,\delta) : (0,1) \times (0,1) \to \mathbb{N}$, כך שלכל $\varepsilon,\delta \in (0,1)$, ולכל התפלגות P, בהנתן מדגם S, עבורו $|S| > N(\varepsilon,\delta)$, מתקיים –

$$\mathbb{P}\left[e_P\big(A(S)\big) > \varepsilon\right] \le \delta$$

נרצה להראות כי $\mathcal{H}$ היא $PAC\ learnable$ **בתוחלת**, ולשם כך נדרש להוכיח כי קיים אלגוריתם לומד A^*, ופונק' $N^*(a)$, שמוגדרת ע"י $N^*(a) : (0,1) \to \mathbb{N}$, כך שלכל $a \in (0,1)$, ולכל התפלגות P, בהנתן מדגם S, עבורו $|S| > N^*(a)$, מתקיים –

$$\mathbb{E}\left[e_P\big(A(S)\big)\right] \le a$$

ואכן, נוכיח כי עבור $A^* := A$, כלומר **<u>אלגוריתם זהה</u>, ועבור** $\boldsymbol{N^*(a)}$ **המוגדרת ע"י –** $\boldsymbol{N^*(a) := N\left(\frac{a}{2},\frac{a}{2}\right)}$, מקיימת את התנאי הדרוש.

יהי $a \in (0,1)$, תהי התפלגות P, ויהי מדגם S, עבורו $|S| > N^*(a)$. **נוכיח** כי מתקיים –

$$\mathbb{E}\left[e_P\big(A(S)\big)\right] \le a$$

כיוון שהנחנו $|S| > N^*(a) = N\left(\frac{a}{2},\frac{a}{2}\right)$, ניתן להסיק מכך שהמחלקה $\mathcal{H}$ היא $PAC\ learnable$,

עבור $\varepsilon = \frac{a}{2},\ \ \delta = \frac{a}{2}$, נקבל –

$$\mathbb{P}\left[\; e_P\big(A(S)\big) > \underbrace{\frac{a}{2}}_{=\varepsilon} \;\right] \le \underbrace{\frac{a}{2}}_{=\delta}$$

על ידי שימוש במשלים של ההסתברות הזו, נקבל –

$$\mathbb{P}\left[e_P\big(A(S)\big) > \frac{a}{2}\right] = \; 1 \; - \; \mathbb{P}\left[e_P\big(A(S)\big) \le \frac{a}{2}\right]$$

ולכן נקבל מכאן –

$$1 - \frac{a}{2} \le \mathbb{P}\left[e_P\big(A(S)\big) \le \frac{a}{2}\right] \le 1$$

כמו כן, ניתן להניח כי פונקציית ההפסד חסומה בין 0 ל-1, כלומר –

$$0 \le \Delta(A(S)(X), Y) \le 1$$

ומכאן ניתן להסיק כי –

$$e_P\big(A(S)\big) = \mathbb{E}_P[\Delta(A(S)(X), Y)] \le \mathbb{E}_P[1] = 1$$

נשתמש בנוסחת התוחלת השלמה כדי לקבל –

$$\mathbb{E}\big[e_P\big(A(S)\big)\big] = \overbrace{\mathbb{E}\left[e_P\big(A(S)\big) \mid e_P\big(A(S)\big) > \frac{a}{2}\right]}^{\le 1} \cdot \overbrace{\mathbb{P}\left[e_P\big(A(S)\big) > \frac{a}{2}\right]}^{\le \frac{a}{2}} +$$
$$+ \underbrace{\mathbb{E}\left[e_P\big(A(S)\big) \mid e_P\big(A(S)\big) \le \frac{a}{2}\right]}_{\le \frac{a}{2}} \cdot \underbrace{\mathbb{P}\left[e_P\big(A(S)\big) \le \frac{a}{2}\right]}_{\le 1} = \frac{a}{2} + \frac{a}{2} = a$$

וכאמור קיבלנו –

$$\mathbb{E}\big[e_P\big(A(S)\big)\big] \le a$$

ובעצם הוכחנו כי המחלקה $\mathcal{H}$ היא $PAC\ learnable$ בתוחלת, לפי ההגדרה, כנדרש.

שאלה 3. Union of Intervals.

פיתרון. $\mathcal{H}_k$ מוגדרת על ידי –

$$\mathcal{H}_k = \{h_I \mid I = \{[\ell_1, u_1], \ldots, [\ell_k, u_k], \ \ell_1 \le u_1 \le \ell_2 \le \cdots \le \ell_k \le u_k\}\}$$

כאשר $h_I(x) = 1$ אם ורק אם x שייך לאחד מהקטעים ב-I.

נוכיח כי $VC - dimension$ של $\mathcal{H}_k$ הינו $2k$. לשם כך, נדרש להוכיח שני תנאים.

ראשית, **נראה כי $VC - dimension$ של $\mathcal{H}_k$ הינו לכל הפחות $2k$.** תהי קבוצה S בגודל $2k$ של נקודות על הישר, ונניח כי S ממוינת(ללא הגבלת הכלליות). נסמן –

$$S = \{x_1, x_2, \ldots, x_{2k} \mid x_1 < x_2 \ldots < x_{2k}\}$$

הטענה המרכזית היא שניתן ליצור כל דיכוטומיה $[s_1, \ldots, s_{2k}]$ על ידי שימוש ב-k-אינטרוולים. הטענה נובעת מכך שניתן להשתמש באינטרוול עבור **כל רצף** של $s_i, \ldots, s_j$ של תיוגים חיוביים($positive\ labeling$), כאשר הכוונה בשימוש באינטרוול עבור רצף הינו הגדרתו בקטע המכיל את כל הנקודות $x_i, \ldots, x_j$, למשל ע"י $[x_i, x_j]$.

כך, כל s_i עם תיוג חיובי, נמצא ברצף כלשהו של תיוגים חיוביים, ולכן הנקודה המתאימה x_i מוכלת באינטרוול כלשהו, מהגדרתנו. מצד שני, כל s_i עם תיוג שלילי, אינה נמצאת ברצף כלשהו של תיוגים חיוביים, ולכן מהגדרתנו של האינטרוולים הללו, הנקודה המתאימה x_i אינה מוכלת באף אינטרוול.

ובכל מקרה, על ידי הגדרה של האינטרוולים באופן זה, נקבל $h_I \in \mathcal{H}_k$, כך שמתקיים $h_I(x_i) = s_i$, כנדרש.

(נעיר כי אם יש פחות מ-k רצפים של תיוגים חיוביים, ניתן להגדיר את האינטרוולים הנותרים כאחרון המוגדר.)

כעת, על מנת להשלים את הנכונות, <u>נוכיח כי ישנם לכל היותר k רצפים של תיוגים חיוביים</u>, ולכן נדרשים לכל היותר k אינטרוולים על מנת להתאים $h_I \in \mathcal{H}_k$. ואכן, נוכיח זאת בשלילה, ונניח כי ישנם לפחות $k + 1$ רצפים של תיוגים חיוביים. באופן טבעי, נסיק לכן שיש ישנם לפחות k רצפים של תיוגים שלילים בין הרצפים החיוביים, שכן אחרת, אם בין זוג רצפים חיוביים לא היה רצף של תיוגיים שליליים, היה זה למעשה אותו הרצף החיובי, בסתירה. על כן, מכך שכל רצף מכיל לפחות תיוג אחד, כלומר בכל רצף לכל הפחות s_i אחד, ומכך שיש לפחות $2k + 1$ רצפים (כלשהם), נסיק כי ישנם לפחות $2k + 1$ איברים בדיכוטומיה, **בסתירה** לכך שהנחנו כי בדיכוטומיה ישנם בדיוק $2k$ איברים, שהם $[s_1, \ldots, s_{2k}]$. על כן, **ניתן להסיק כאמור כי ישנם לכל היותר k רצפים של תיוגים חיוביים, ולכן נדרשים לכל היותר k אינטרוולים, כנדרש מהשלמת נכונות הטענה.**

בסה"כ, הראנו עד כה, כי $VC - dimension$ של $\mathcal{H}_k$ הינו לכל הפחות $2k$.

נותר אם כן להראות כי $VC - dimension$ של $\mathcal{H}_k$ הינו לכל היותר $2k$. תהי קבוצה S בגודל $2k + 1$ של נקודות על הישר, ונניח כי S ממוינת(ללא הגבלת הכלליות). נסמן $S = \{x_1, \ldots, x_{2k+1} \mid x_1 < \ldots < x_{2k+1}\}$.

נרצה להראות כי $|\mathcal{H}_{k_S}| < 2^{|S|} = 2^{2k+1}$, ולשם כך, נראה דיכוטומיה בגודל 2k+1, שנסמן $[s_1, \ldots, s_{2k+1}]$ שעבורה <u>לא קיימת</u> $h_I \in \mathcal{H}_k$ כך שמתקיים $h_I(x_i) = s_i$. נגדיר כך –

$$s_i = 1 \quad for \quad i \in \mathbb{N}_{odd} \quad , \ and \ s_i = 0 \quad for \quad i \in \mathbb{N}_{even}$$

כך שניתן לכתוב בצורה מפורשת - $[s_1, \ldots, s_{2k+1}] = [1, 0, 1, \ldots, 0, 1]$. נוכיח בשלילה, ונניח כי ישנה $h_I \in \mathcal{H}_k$ כך שמתקיים $h_I(x_i) = s_i$. נסמן ב- $[\ell_i, u_i]$ את האינטרוול ה-i ב— I. הסתירה כאן נובעת מכך שיש ב-I לפחות $k + 1$ אינטרוולים, בסתירה להגדרת $\mathcal{H}_k$. ואכן, יש בדיוק $k + 1$ רצפים של תיוגים חיוביים, שהינם במקרה זה $x_1, x_3, x_5, \ldots, x_{2k+1}$, שכן כאמור לכל $1 \le i \le k$, מתקיים ש- $s_{2i} = 0$, ולכן אף אינטרוול אינו מכיל את $x_2, x_4, \ldots, x_{2k}$, **ועל כן נדרשים בדיוק $k + 1$ אינטרוולים עבור דיכוטומיה זו**, **וזו סתירה כאמור לכך שקיימת $h_I \in \mathcal{H}_k$ עבורה מתקיים $h_I(x_i) = s_i$.**

מכאן, ניתן להסיק כי $VC - dimension$ של $\mathcal{H}_k$ הינו לכל היותר $2k$.

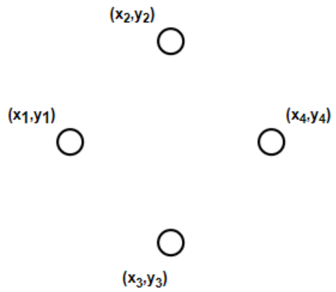
ובסה"כ, הראנו כי ה— $VC - dimension$ של $\mathcal{H}_k$ הינו בדיוק $2k$.

שאלה 4. Right-Angle Triangles.

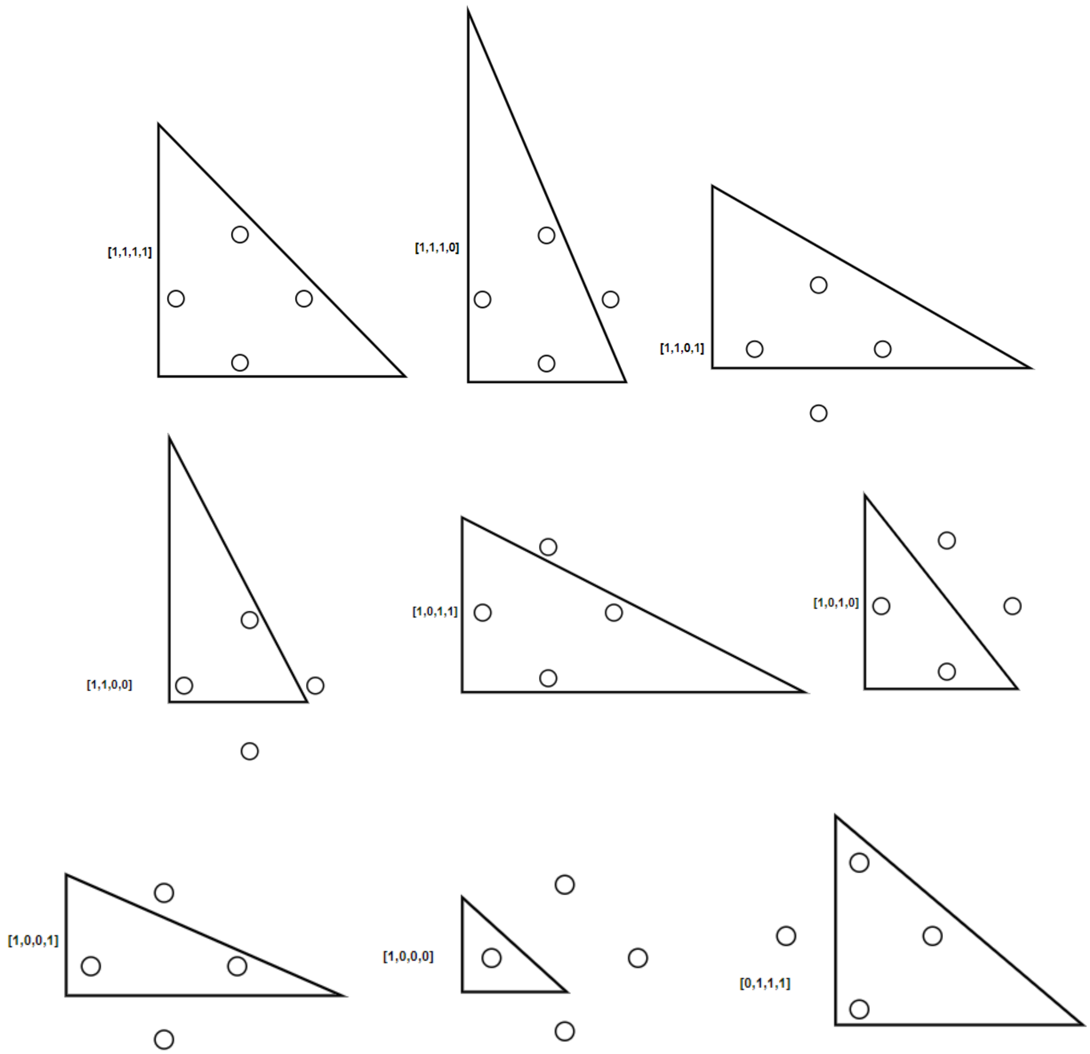
נמצא ונוכיח את ה-$VC - dimension$ של המחלקה הבאה –

$$\boldsymbol{\mathcal{H}\ is\ defined\ as\ the\ class\ of\ hypotheses\ of\ axis - aligned\ right\ angle\ triangle}$$
$$\boldsymbol{in\ the\ plane, with\ the\ right\ angle\ in\ the\ lower\ left\ corner}.$$

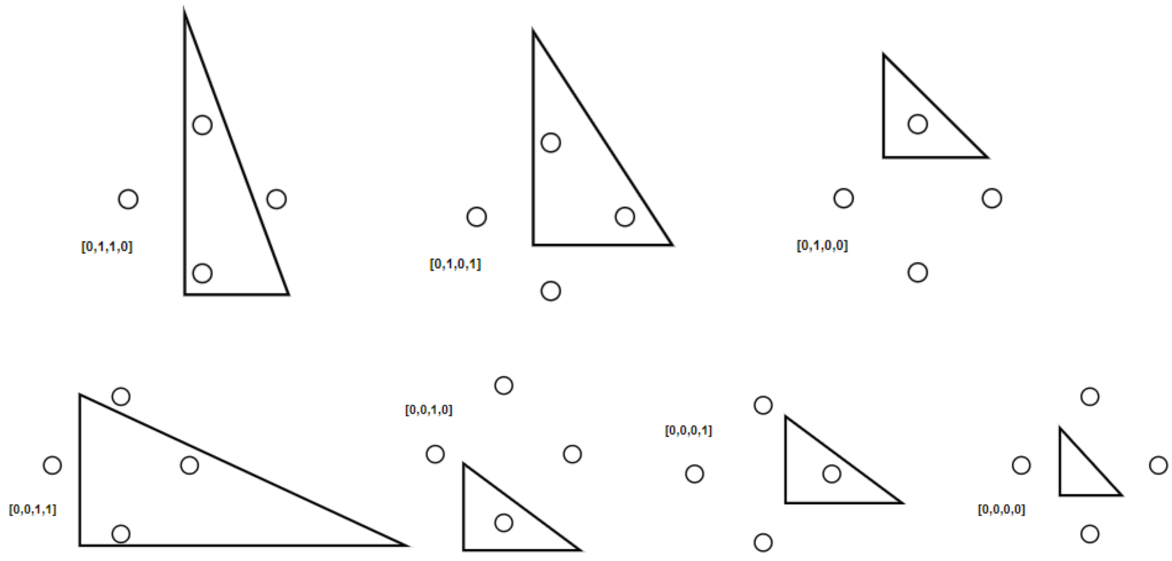
פיתרון. נוכיח כי ה-$VC - dimension$ של $\mathcal{H}$ הינה 4, כלומר $\boldsymbol{VC - dim}(\mathcal{H}) = 4$. לשם כך נדרש להראות את שני התנאים הבאים.
ראשית נוכיח כי $\boldsymbol{VC - dim}(\mathcal{H}) \geq 4$ על ידי כך שנראה קבוצה $\boldsymbol{S} \subseteq \mathbb{R}^2$ מגודל 4, שמקיימת $|\mathcal{H}_S| = 2^{|S|}$.
נבחר את הקבוצה S באופן הבא, ונסמן – $S = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\}$ כך שיתקיימו כמה תכונות: $x_1 < x_4$ וכן $y_1 = y_4$, ובנוסף $x_2 = x_3$ וכן $y_2 > y_3$, ובאופן סכמתי ניתן לחשוב על הנק' באופן הבא –



נראה את השיוויון $|\mathcal{H}_S| = 2^{|S|} = 16$, על ידי כך שנראה כי כל דיכוטומיה יכולה להתקבל על ידי $h \in \mathcal{H}$,
כלומר נראה שלכל דיכוטומיה $[s_1, s_2, s_3, s_4]$ קיימת $h \in \mathcal{H}$ כך ש-$h\big((x_i, y_i)\big) = s_i$, לכל $1 \leq i \leq 4$.
אופן ההוכחה יהיה למעשה הוכחה ע"י תמונה($proof\ by\ picture$), כך שנראה שלכל דיכוטומיה,
כלומר לכל תת-קבוצה של נקודות(תת הקבוצה נקבעת לפי $s_i = 1$), קיים משולש ישר זווית(שמאלי)
המקביל למישור x כך שהמשולש מכיל בדיוק את הנקודות בתת-הקבוצה הזו(ובפרט, לא מכיל את
הנקודות שאינן בתת-הקבוצה הזו). כעת, נדרש להראות זאת לכל דיכוטומיה אפשרית –



המשך שאלה 4.



על כן, זה מוכיח את השיוויון $|\mathcal{H}_S| = 2^{|S|} = 16$, שכן הראנו כי **הקבוצה** S **הינה** $\boldsymbol{shattered\ by\ \mathcal{H}}$**, ולכן ניתן להסיק את האי-שיוויון בכיוון הראשון –** $\boldsymbol{VC - dim(\mathcal{H}) \geq 4}$**.**

נותר להוכיח את האי-שיוויון בכיוון השני - $\boldsymbol{VC - dim(\mathcal{H}) \leq 4}$, ובפרט נראה שלא קיימת קבוצת מדגם S, כך ש- $\boldsymbol{|S| \geq 5}$, עבורה מתקיים – $|\mathcal{H}_S| = 2^{|S|}$. ראינו בהרצאה כי מספיק להוכיח זאת עבור קבוצה מגודל מינימלי, כלומר מקבוצה בגודל **5**, במקרה זה. תהי קבוצה מגודל 5, שנסמן ב- $\boldsymbol{S = \{x_1, x_2, \ldots, x_5\} \subseteq \mathbb{R}^2}$. זו קבוצה של 5 נקודות במישור.
נתבסס על שיקולים גיאומטריים בסיסיים, ונחלק את ההוכחה לשני מקרים אפשריים.
מקרה 1. קיימת נקודה שנמצאת בתוך הקמור($Convex\ hull$) של הנקודות. במקרה זה, נסמן ב- x_ℓ, x_r את הנקודות ב-S, עבורן x_ℓ השמאלית ביותר, x_r הימנית ביותר, ובן נסמן גם כן x_{high}, x_{low} את הנקודות ב-S, עבורן x_{low} הנמוכה ביותר, ו-x_{high} הגבוהה ביותר מביניהן. מדובר בלכל היותר 4 נקודות שונות, ולכן קיימת נקודה $x_j \in S$, כך שמתקיים – $x_j \neq x_\ell, x_r, x_{high}, x_{low}$, כלומר ניתן לומר כי הנקודה x_j נמצאת בין כל שאר הנקודות. הטענה המרכזית והטבעית היא שלא קיים משולש במישור המכיל את כל הנקודות $x_\ell, x_r, x_{high}, x_{low}$ ולא מכיל את הנקודה x_j, כאשר ההוכחה מיידית ונובעת משיקולים גיאומטריים בסיסיים. מכאן, נבחין כי הדיכוטומיה $[s_1, s_2, \ldots, s_5]$ המוגדרת ע"י $s_i = 0 \iff i = j$, מייצגת למעשה את המקרה הנ"ל, שבו נרצה משולש שיכיל את כל הנקודות מלבד x_j, ומנכונות הטענה הגיאומטרית שצוינה, נסיק כי לא קיימת $h \in \mathcal{H}$, כך שמתקיים השיוויון $h(x_i) = s_i$, ובפרט נסיק כי במקרה זה הקבוצה S אינה $shattered\ by\ \mathcal{H}$, ומכך ש- S הינה קבוצה כללית(תחת הנחת מקרה 1), נסיק כי לא קיימת תת-קבוצה של $\mathbb{R}^2$, מגודל 5, שהיא $shattered\ by\ \mathcal{H}$, מה שגורר כי $VC - dim(\mathcal{H}) \leq 4$.
מקרה 2. כל הנקודות ב-S נמצאות על הקמור של הנקודות, ובמקרה זה ההוכחה הינה דומה ומכילה שיקולים גיאומטריים דומים. גם במקרה זה נסיק כי לא קיימת תת-קבוצה S(תחת הנחת מקרה 2) של $\mathbb{R}^2$, מגודל 5, שהיא $shattered\ by\ \mathcal{H}$, מה שגורר $VC - dim(\mathcal{H}) \leq 4$.

ובכל מקרה, קיבלנו כי $\boldsymbol{VC - dim(\mathcal{H}) \leq 4}$.

ובסה"כ, הוכחנו את כל התנאים הנדרשים, ולכן נסיק כי $\boldsymbol{VC - dimension}$ **של המחלקה הינו 4 בדיוק, כנדרש.**

שאלה 5. Structural Risk Minimization.

סעיף א.

נתונה מחלקה $\mathcal{H} = \cup_{i\in\mathbb{N}}\{h_i\} = \cup_{i\in\mathbb{N}}\mathcal{H}_i$, כאשר נסמן למעשה $\mathcal{H}_i = \{h_i\}$. נרצה להוכיח כי בהסתברות של לפחות $1-\delta$, מתקיים –

$$\forall h \in \mathcal{H}.\, e_P(h) \le e_S(h) + \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h)}\right)}$$

ראינו את ה – $uniform\ convergence\ bound$ עבור מחלקות סופיות ולפיו –

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|e_S(h) - e_P(h)| \ge \varepsilon\right] \le 2|\mathcal{H}|e^{-2n\varepsilon^2}$$

נשתמש בחסם זה עבור כל מחלקה $\mathcal{H}_i$, ונרצה חסם הסתברותי של $\delta_i = \delta \cdot w(h_i)$, וכיוון ש – $|\mathcal{H}_i| = 1$ לכל $i \in \mathbb{N}$, נקבל –

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}_i}|e_S(h) - e_P(h)| \ge \varepsilon\right] = \mathbb{P}[|e_S(h_i) - e_P(h_i)| \ge \varepsilon] \le 2|\mathcal{H}_i|e^{-2n\varepsilon^2} = 2e^{-2n\varepsilon^2} \underbrace{\le}_{!} \delta_i = \delta \cdot w(h_i)$$

ונקבל כי מתקיים –

$$2e^{-2n\varepsilon^2} \le \delta \cdot w(h_i) \quad \Leftrightarrow \quad \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h_i)}\right)} \le \varepsilon$$

כלומר, קיבלנו כי –

$$\mathbb{P}\left[|e_S(h_i) - e_P(h_i)| \ge \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h_i)}\right)}\right] \le \delta_i$$

וזה נכון לכל $i \in \mathbb{N}$. כעת נסתכל על המחלקה $\mathcal{H}$ כולה, ונקבל עבורה ע"י חסם האיחוד –

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|e_S(h) - e_P(h)| \ge \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h)}\right)}\right] = \mathbb{P}\left[\exists h \in \mathcal{H}\ \ s.t.\ \ |e_S(h) - e_P(h)| \ge \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h)}\right)}\right] \le$$

$$\le \sum_{i=0}^{\infty} \mathbb{P}\left[|e_S(h_i) - e_P(h_i)| \ge \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h_i)}\right)}\right] \le \sum_{i=0}^{\infty}\delta_i = \sum_{i=0}^{\infty}\delta \cdot w(h_i) = \delta \cdot \sum_{i=0}^{\infty} w(h_i) \le \delta$$

כאשר נזכור כי $\sum_{i=0}^{\infty} w(h_i) \le 1$ לפי הנתון. מכאן נקבל –

$$\mathbb{P}\left[\forall h \in \mathcal{H}.\ \ |e_S(h) - e_P(h)| \le \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h)}\right)}\right] \ge \mathbb{P}\left[\sup_{h\in\mathcal{H}}|e_S(h) - e_P(h)| \le \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h)}\right)}\right] =$$

$$= 1 - \mathbb{P}\left[\sup_{h\in\mathcal{H}}|e_S(h) - e_P(h)| \ge \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h)}\right)}\right] \ge 1 - \delta$$

לפיכך, קיבלנו כי בהסתברות של לפחות $1-\delta$, מתקיים –

$$\forall h \in \mathcal{H}. \quad e_P(h) \le e_S(h) + \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta \cdot w(h)}\right)}$$

כנדרש.

סעיף ב.

נגדיר $h_{srm} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, e_S(h) + \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h)}\right)}$, וכן נגדיר גם – $h_{min} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, e_P(h)$.
בסעיף א' הוכחנו כי בהסתברות של לפחות $1-\delta$, מתקיים –

$$\forall h \in \mathcal{H}. \quad |e_S(h) - e_P(h)| \le \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h)}\right)}$$

כלומר בהסתברות של לפחות $1-\delta$, מתקיים –

$$\forall h \in \mathcal{H}. \quad -\sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h)}\right)} \le e_S(h) - e_P(h) \le \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h)}\right)}$$

כאשר נבחין כי הנ"ל יתקיים <u>לכל</u> $h \in \mathcal{H}$ בהסתברות זו. מתקיים –

$$e_P(h_{min}) \underset{\substack{h_{min}\\ minimizer\\ for\ e_P}}{\leqq} e_P(h_{srm}) \underset{\substack{w.p.\\ \ge 1-\delta}}{\leqq} e_S(h_{srm}) + \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h_{srm})}\right)} \underset{\substack{h_{srm}\\ minimizer\\ for\ e_S+penalty}}{\leqq} e_S(h_{min}) + \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h_{min})}\right)}$$

כלומר קיבלנו כי בהסתברות של לפחות $1-\delta$, מתקיים –

$$e_P(h_{srm}) - e_S(h_{min}) \le \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h_{min})}\right)}$$

וגם

$$e_S(h_{min}) - e_P(h_{min}) \le \sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h_{min})}\right)}$$

כמו כן, מתקיים – $e_P(h_{min}) \le e_P(h_{srm})$, שכן h_{min} מביא למינ' את e_P על הקבוצה $\mathcal{H}$. לכן, בהסתברות של לפחות $1-\delta$, מתקיים –

$$e_P(h_{srm}) - e_P(h_{min}) = \underbrace{e_P(h_{srm}) - e_S(h_{min})} + \underbrace{e_S(h_{min}) - e_P(h_{min})} \le 2\cdot\sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h_{min})}\right)}$$

וביוון ש – $0 \le e_P(h_{srm}) - e_P(h_{min})$, נקבל כי $|e_P(h_{srm}) - e_P(h_{min})| = e_P(h_{srm}) - e_P(h_{min})$, ומכאן נסיק,
כי בהסתברות של לפחות $1-\delta$, מתקיים –

$$|e_P(h_{srm}) - e_P(h_{min})| \le 2\cdot\sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h_{min})}\right)}$$

וביוון שמתקיים – $e_P(h_{min}) = \min_{h\in\mathcal{H}} e_P(h)$ בעצם לפי ההגדרה, שכן $h_{min} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, e_P(h)$, נקבל את הנדרש –

בהסתברות של לפחות $1-\delta$, מתקיים –

$$\left|e_P(h_{srm}) - \min_{h\in\mathcal{H}} e_P(h)\right| \le 2\cdot\sqrt{\frac{1}{2n} ln\left(\frac{2}{\delta\cdot w(h_{min})}\right)}$$

כנדרש.

שאלה 6. Loss Minimization.

נרצה למצוא מסווג אופטימלי($optimal\ classifier$). יהי $x_0 \in \mathcal{X}$ כלשהו, ונרצה להבין מה הוא הערך האופטימלי עבור $h(x_0)$ כך שיתקבל –

$$h = \underset{h\in\mathcal{H}}{argmin}\, \mathbb{E}[\Delta_b(Y, h(X)]$$

במילים אחרות, נרצה למצוא את הערך האופטימלי עבור $h(x_0)$, כך שנקבל ש- h הינו $optimal\ classifier$.

מתקיים –

$$\mathbb{E}[\Delta_b(Y, h(X)] = \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \mathbb{P}[Y = y, X = x] \cdot \Delta_b(y, h(x)) =$$

$$= \underbrace{\sum_{\substack{x\in\mathcal{X}\setminus\{x_0\}\\ y\in\mathcal{Y}}} \mathbb{P}[Y = y, X = x] \cdot \Delta_b\big(y,\ \ h(x)\big)}_{doesn't\ \ depend\ \ on\ \ x_0} + \underbrace{\sum_{y\in\mathcal{Y}} \mathbb{P}[Y = y, X = x_0] \cdot \Delta_b\big(y,\ \ h(x_0)\big)}_{depends\ \ on\ \ x_0\ \ and\ \ h(x_0)}$$

כאמור, נרצה למצוא את הערך האופטימלי, עבור x_0, כך ש-$h(x_0)$ יגרור מינימליות של $\mathbb{E}[\Delta_b(Y, h(X)]$.

על כן, מספיק להסתכל על הגורמים בביטוי זה התלויים ב-x_0, כלומר ב-$h(x_0)$. על כן, נדרש להביא למינימום את הביטוי –

$$\sum_{y\in\mathcal{Y}} \mathbb{P}[Y = y, X = x_0] \cdot \Delta_b\big(y, h(x_0)\big) =$$

$$= \mathbb{P}[Y = 1, X = x_0] \cdot \Delta_b\big(1\ , h(x_0)\big) + \mathbb{P}[Y = 0, X = x_0] \cdot \Delta_b\big(0, h(x_0)\big) =$$

$$= \underbrace{\mathbb{P}[X = x_0]}_{constant} \cdot \Big(\mathbb{P}[Y = 1 \mid X = x_0] \cdot \Delta_b\big(1, h(x_0)\big) + \mathbb{P}[Y = 0 \mid X = x_0] \cdot \Delta_b\big(0, h(x_0)\big)\Big)$$

נסתכל כעת על המקרים האפשריים.

אם $h(x_0) = 0$, נקבל כי הביטוי שקול ל– $\mathbb{P}[X = x_0] \cdot \left(\mathbb{P}[Y = 1 \mid X = x_0] \cdot \frac{1}{2}\right)$.

אם $h(x_0) = 1$, נקבל כי הביטוי שקול ל– $\mathbb{P}[X = x_0] \cdot \mathbb{P}[Y = 0 \mid X = x_0]$.

כמו כן, כיוון שנרצה להביא למינימום את הביטוי הנ"ל, וכיוון שההסתברות הללו ידועות, נסיק כי נדרש לבחור לכל $x \in \mathcal{X}$ –

$$h(x) = \begin{cases} 1 & if\ \ \ \mathbb{P}[Y = 0 \mid X = x] < \mathbb{P}[Y = 1 \mid X = x] \cdot {}^{1}\!/_{2} \\ 0 & else \end{cases}$$

בנוסף, מתקיים כי – $\mathbb{P}[Y = 0 \mid X = x] = 1 - \mathbb{P}[Y = 1 \mid X = x]$, ולכן מתקיים –

$$\mathbb{P}[Y = 0 \mid X = x] < \mathbb{P}[Y = 1 \mid X = x] \cdot \frac{1}{2} \quad iff \quad \mathbb{P}[Y = 1 \mid X = x] > \frac{2}{3}$$

ולכן נקבל את התנאי השקול הבא -

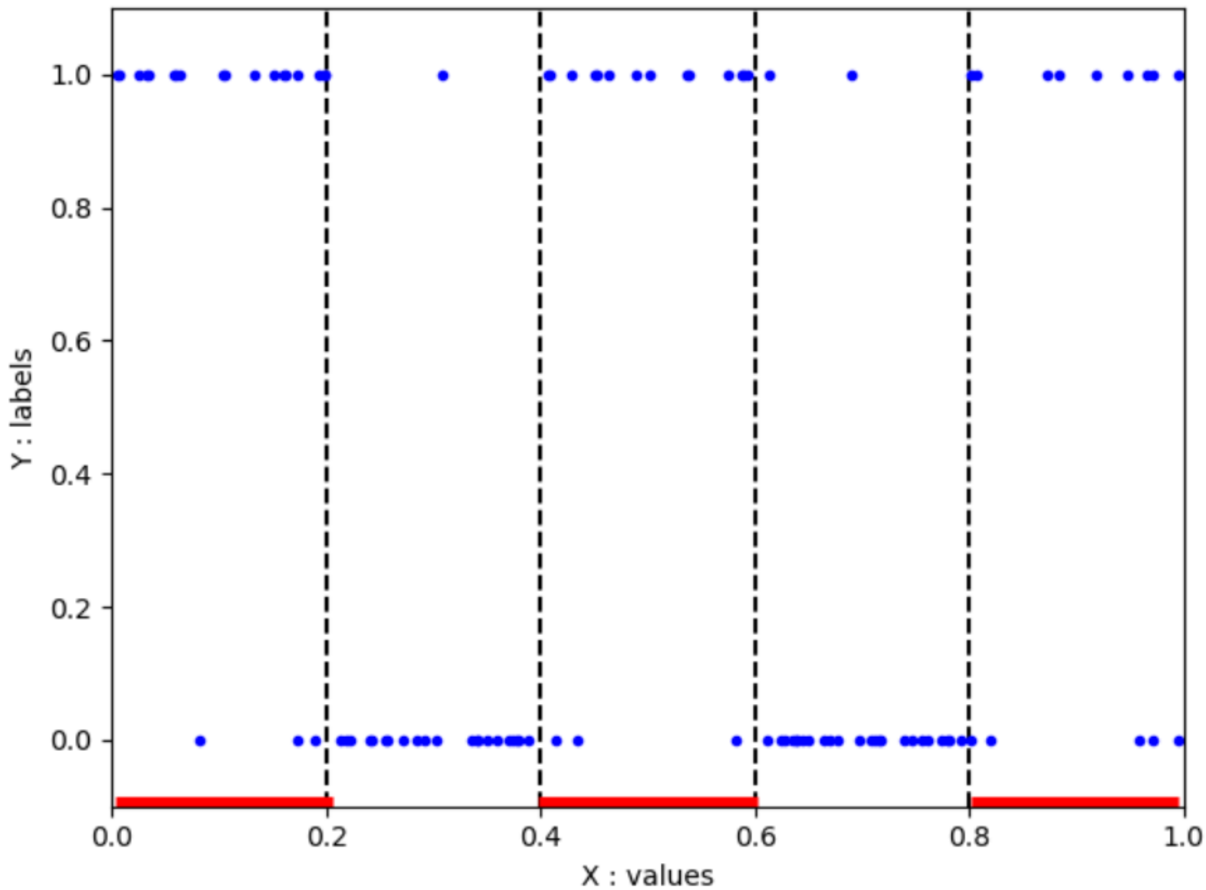
$$h(x) = \begin{cases} 1 & if\ \ \ \mathbb{P}[Y = 1 \mid X = x] > \frac{2}{3} \\ 0 & else \end{cases}$$

כנדרש. **המסווג הבינארי הנ"ל הינו האופטימלי ומביא את $\mathbb{E}[\Delta_b(Y, h(X)]$ למינימום.**

שאלה 1 – Programming Assignment.

סעיף א.

בדוגמה זו דגמנו $m = 100$ נקודות ביחס להתפלגות P, והרצנו את אלגוריתם ה-ERM עם פרמטר $k = 3$.
קיבלנו את התוצאות הבאות: הנקודות הן בכחול, והקטעים שקיבלנו מאלגוריתם ה-ERM הם באדום.



סעיף ב.

כאן $\mathcal{X} = [0,1]$, וכן $\mathcal{Y} = \{0,1\}$. כמו כן, $X \sim U(0,1)$. לכל $h \in \mathcal{H}_k$, מתקיים –

$$error(h) = e_P(h) = \mathbb{E}_P[\Delta_{\text{zo}}(h(X), Y)] =$$
$$= \int_0^1 \left(\sum_{y \in \mathcal{Y}} \mathbb{P}[Y = y, X = x] \cdot \Delta_{\text{zo}}(h(x), y) \right) dx =$$
$$= \int_0^1 \left(\mathbb{P}[Y = 0, X = x] \cdot \Delta_{\text{zo}}(h(x), 0) \ + \ \mathbb{P}[Y = 1, X = x] \cdot \Delta_{\text{zo}}(h(x), 1) \right) dx$$
$$= \int_0^1 \left(\mathbb{P}[Y = 0 \mid X = x] \cdot f_X(x) \cdot \Delta_{\text{zo}}(h(x), 0) \ + \ \mathbb{P}[Y = 1 \mid X = x] \cdot f_X(x) \cdot \Delta_{\text{zo}}(h(x), 1) \right) dx$$
$$= \int_0^1 f_X(x) \cdot \left(\mathbb{P}[Y = 0 \mid X = x] \cdot \Delta_{\text{zo}}(h(x), 0) \ + \ \mathbb{P}[Y = 1 \mid X = x] \cdot \Delta_{\text{zo}}(h(x), 1) \right) dx$$

מהנתונים בשאלה, נזכור כי $\mathbb{P}[Y = 1 \mid X = x]$ תלויה ומוגדרת לפי הקטע בו x נמצא, ובכל אחד מבין הקטעים הללו מדובר בקבוע, וכנ"ל באמור על המשלים – $\mathbb{P}[Y = 0 \mid X = x]$. לפיכך, נרצה להסתכל על האינטגרל הכולל בסכום של אינטגרלים על הקטעים הנתונים. כלומר, נקבל –

$$e_P(h) = \sum_{i=0}^{4} \int_{0.2i}^{0.2i+0.2} f_X(x) \left(\mathbb{P}[Y = 0 | X = x] \cdot \Delta_{\text{zo}}(h(x), 0) + \mathbb{P}[Y = 1 | \ X = x] \cdot \Delta_{\text{zo}}(h(x), 1) \right) dx$$

כאמור, המטרה הינה למצוא $h \in \mathcal{H}_k$ עבורה מתקיים ש – $e_P(h)$ מינימלית.
עבור הקטעים $[0.8, 1]$, $[0.4, 0.6]$, $[0, 0.2]$, נתון $\mathbb{P}[Y = 1 | X = x] = 0.8$, ולכן $\mathbb{P}[Y = 0 | X = x] = 0.2$.
עבור הקטעים $[0.6, 0.8]$, $[0.2, 0.4]$, נתון כי $\mathbb{P}[Y = 1 \mid x] = 0.1$, ולכן $\mathbb{P}[Y = 0 \mid x] = 0.9$.

עבור $\boldsymbol{i = 0, 2, 4}$, נקבל את הקטעים $[0.8, 1]$, $[0.4, 0.6]$, $[0, 0.2]$ וכן הגורמים המתאימים יהיו –

$$\int_{0.2i}^{0.2i+0.2} f_X(x) \left(0.2 \cdot \Delta_{\text{zo}}(h(x), 0) + 0.8 \cdot \Delta_{\text{zo}}(h(x), 1) \right) dx$$

ובקטעים אלה מתקיים –

$$\underbrace{\int_{0.2i}^{0.2i+0.2} 0.2 \cdot f_X(x) dx}_{if \quad h(x)=1} \ \le \int_{0.2i}^{0.2i+0.2} f_X(x) \left(0.2 \cdot \Delta_{\text{zo}}(h(x), 0) + 0.8 \cdot \Delta_{\text{zo}}(h(x), 1) \right) dx \le \underbrace{\int_{0.2i}^{0.2i+0.2} 0.8 \cdot f_X(x) dx}_{if \quad h(x)=0}$$

עבור $\boldsymbol{i = 1, 3}$, נקבל את הקטעים $[0.6, 0.8]$, $[0.2, 0.4]$ וכן הגורמים המתאימים יהיו –

$$\int_{0.2i}^{0.2i+0.2} f_X(x) \left(0.9 \cdot \Delta_{\text{zo}}(h(x), 0) + 0.1 \cdot \Delta_{\text{zo}}(h(x), 1) \right) dx$$

ובקטעים אלה מתקיים –

$$\underbrace{\int_{0.2i}^{0.2i+0.2} 0.1 \cdot f_X(x) dx}_{if \quad h(x)=0} \ \le \int_{0.2i}^{0.2i+0.2} f_X(x) \left(0.2 \cdot \Delta_{\text{zo}}(h(x), 0) + 0.8 \cdot \Delta_{\text{zo}}(h(x), 1) \right) dx \le \underbrace{\int_{0.2i}^{0.2i+0.2} 0.9 \cdot f_X(x) dx}_{if \quad h(x)=1}$$

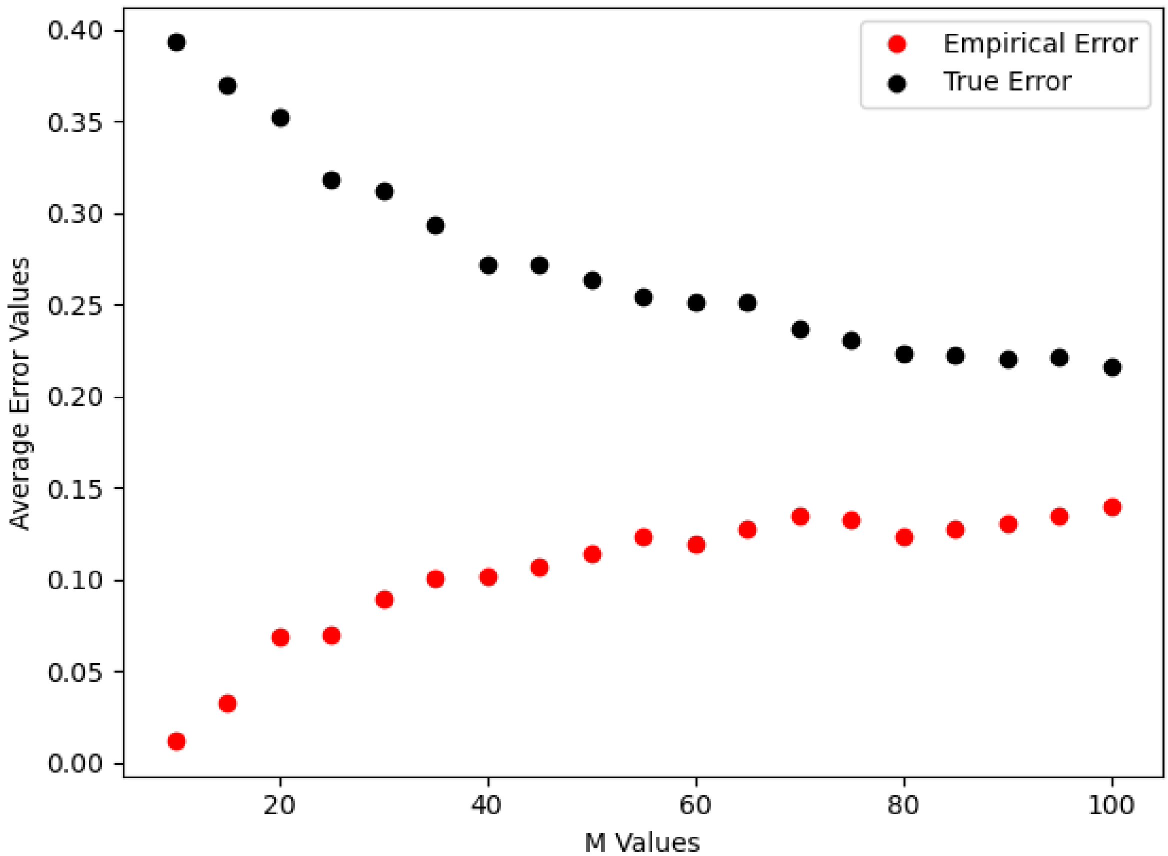
ולכן נרצה להגדיר את $\boldsymbol{h}$ באופן הבא –

$$\boxed{\boldsymbol{h(x) \ = \begin{cases} 1 & if \ x \ \in [0, 0.2] \ \cup \ [0.4, 0.6] \ \cup [0.8, 1] \\ 0 & if \ x \ \in [0.2, 0.4] \ \cup \ [0.6, 0.8] \end{cases}}}$$

על מנת לקבל את השגיאה המינימלית.

סעיף ג.

עבור $k=3$, ועבור $m=10,15,20,\dots,100$, נריץ את אלגוריתם ה-ERM, במשך $T=100$ פעמים, ונמצע את הטעויות האמפיריות והאמיתיות, כדי לקבל את הגרף הבא –



ניתן להבחין כי ישנן שלוש מגמות –

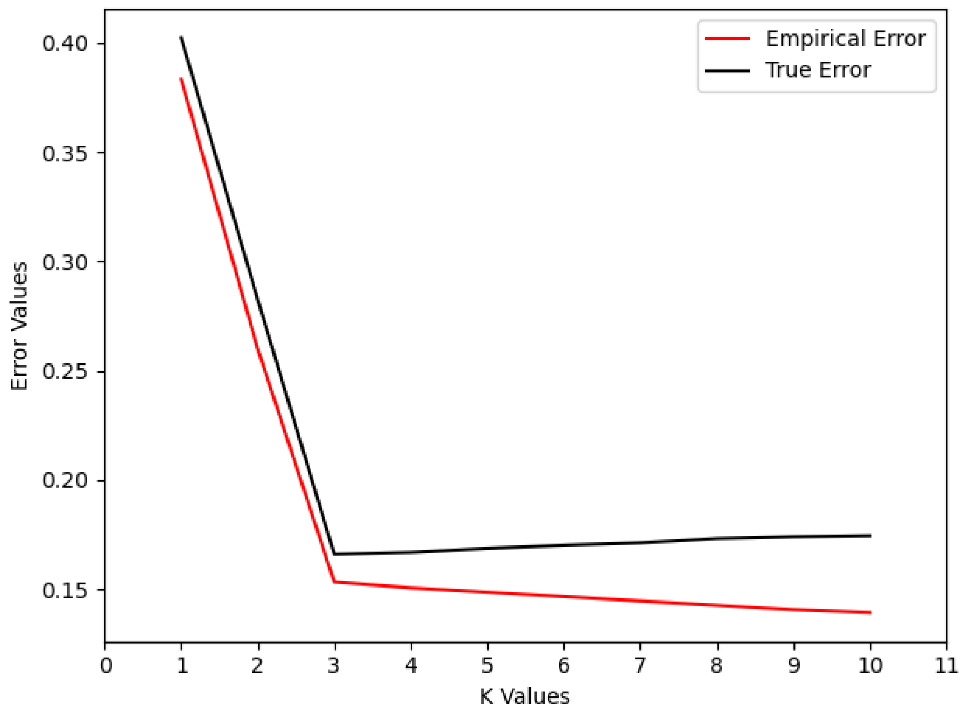
- הטעות האמיתית ($True\ Error$) יורדת ככל שהפרמטר m עולה. הפרמטר m מייצג את מספר הדגימות ב-S, וככל שנראה מספר גדול יותר של דגימות מקריות, המייצגות את ההתפלגות האמיתית, ולכן תהליך הלמידה יפיק היפותזה קרובה יותר להתפלגות האמיתית, ולכן השגיאה האמיתית תקטן ככל ש- m גדל.
במילים אחרות, ככל ש-m גדל, המדגם הנתון גדול יותר, ובפרט מייצג בצורה טובה יותר את ההתפלגות האמיתית, וכיוון שמבצעים אלגוריתם ERM, ובו מביאים למינימום את השגיאה על המדגם, ולכן נקבל כי השגיאה האמיתית קטנה שכן היא מיוצגת טוב יותר ע"י השגיאה האמפירית ככל ש- m גדל.
- הטעות האמפירית ($Empiric\ Error$) עולה ככל שהפרמטר m עולה. כיוון שבמקרה זה, מספר האינטרוולים האפשריים הינו קבוע $k=3$, **נתקשה להתאים שלושה אינטרוולים(לכל היותר) עבור מספר עולה של נקודות במדגם**, ולכן נקבל יותר טעויות על המדגם(ובפרט, נקבל יותר טעויות באופן יחסי לגודל המדגם), ולכן הטעות האמפירית עולה ככל ש m – עולה. במילים אחרות, הטעות האמפירית עולה ככל שגודל המדגם גדל **מכיוון שלאלגוריתם ה-ERM קשה יותר למצוא היפותזה**(המכילה לכל היותר שלושה קטעים) **שתתאים למדגם**, ולכן אחוז השגיאות על המדגם יעלה, ולפיכך הטעות האמפירית עולה ככל שהפרמטר m עולה.
- ראינו בהרצאה כי ישנה התכנסות –

$$\mathbb{P}[\sup_{h\in\mathcal{H}_k}|e_S(h)-e_P(h)|\geq\ \varepsilon\ ]\ \leq 4\Pi_H(2m)e^{\frac{-m\varepsilon^2}{8}}\leq 4\ \left(\frac{em}{2k}\right)^{2k}\ e^{\frac{-m\varepsilon^2}{8}}$$

ונבחין כי החסם מתקרב ל-אפס כאשר m גדל, שכן ישנה דעיכה אקספוננציאלית(למרות שקיים הגורם m^{2k} בחסם), כלומר $e_S(h)$ ו- $e_P(h)$ מתקרבים לכל $h\in\mathcal{H}_k$ כאשר m גדל. ואכן רואים בגרף שכאשר m גדל, הטעות האמפירית והטעות האמיתית מתקרבות זו לזו, כמצופה לפי החסם הנ"ל.

סעיף ד.

נריץ את אלגוריתם ה-ERM, עם $m = 1500$ דגימות, עבור $k = 1,2, \dots ,10$, ונקבל את הגרף הבא –



נבחין כי עבור $k \le 3$, הטעות האמפירית והטעות האמיתית יורדות כפונק' של k, ועבור $k \ge 4$, הטעות האמפירית יורדת כפונק' של k, ואילו הטעות האמיתית עולה כפונק' של k. מסעיף ב', ידוע כי הטעות האמיתית האופטימלית מתקבלת עבור $k = 3$ אינטרוולים, ואכן ניתן להבחין בגרף זה שעבור $k = 3$, מתקבלת הטעות האמיתית המינימלית.

כמו כן, k^* מוגדרת כ-k עבורו מתקבלת הטעות האמפירית המינימלית, וכיוון שהטעות האמפירית הינה פונק' יורדת של k, נקבל $k^* = 10$. באלגוריתם ה-ERM בוחרים את ההיפותזה כך שהיא תמזער את הטעות האמפירית, ולכן בסעיף זה היפותזה זו מתקבלת עבור 10 קטעים.

כמו כן, **נבחין שעבור $k^* = 10$ מתקבלת טעות אמיתית(true error) גדולה יותר מאשר ב– $k = 3$, בו הטעות האמיתית היא מינימלית, ולכן בחירה של היפותזה עם k^* אינטרוולים אינה בהכרח הבחירה האופטימלית**. הסבר אפשרי לכך הינו $overfitting$ אפשרי על גבי המדגם, שאינו בהכרח תואם להתפלגות האמיתית, ולפיכך מתקבלת טעות אמיתית(true error) גדולה יחסית.

סעיף ה.

בסעיף זה נשתמש בעיקרון ה-SRM, על מנת למצוא את ה-k עבורו ה – $test\ error$ טוב ביותר. ראשית, נצטרך לבנות פונק' $penalty$, וזאת נעשה ע"י שימוש בשתי תוצאות שראינו בהרצאה –

$$(\mathbf{1}) \quad \mathbb{P}[\sup_{h\in\mathcal{H}} |e_S(h) - e_P(h)| \ge \ \varepsilon\ ] \ \le 4\Pi_H(2n)e^{\frac{-n\varepsilon^2}{8}}$$

$$(\mathbf{2}) \ \ \Pi_H(n) \le \ \left(\frac{en}{d}\right)^d, as \quad d = VC - dim(\mathcal{H}),\ \ for \quad n > d$$

כמו כן, בשאלה 3 בתרגיל בית זה, ראינו כי $VC - dim(\mathcal{H}_k) = 2k$. ואכן, מתקיים לכן, עבור המחלקה $\mathcal{H}_k$ –

$$\mathbb{P}[\sup_{h\in\mathcal{H}_k} |e_S(h) - e_P(h)| \ge \ \varepsilon\ ] \ \le 4\Pi_H(2n)e^{\frac{-n\varepsilon^2}{8}} \ \le 4\ \left(\frac{en}{2k}\right)^{2k} e^{\frac{-n\varepsilon^2}{8}} \underset{!}{\le} \delta$$

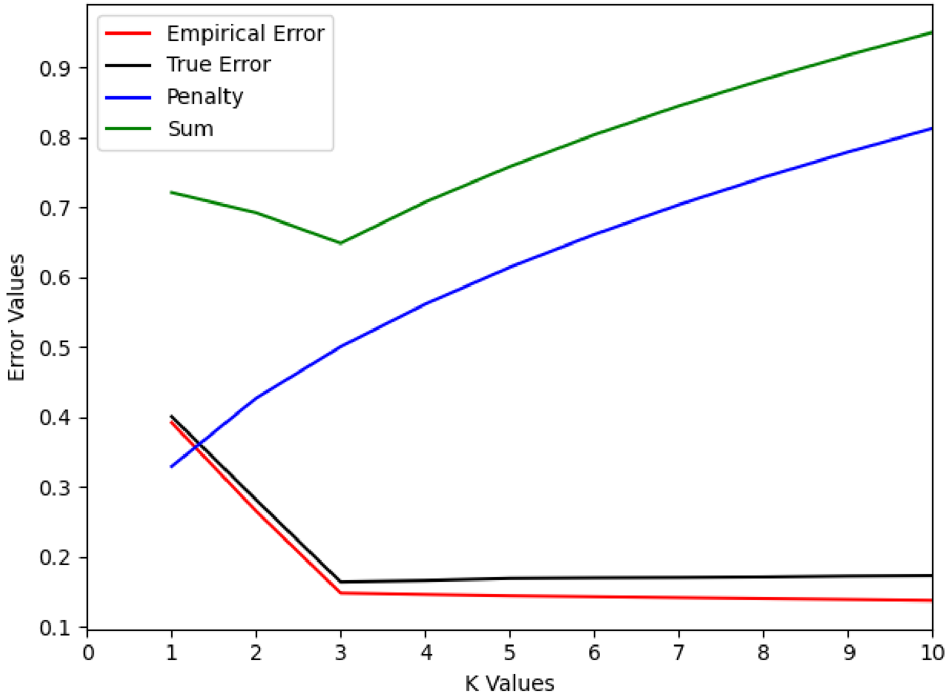
נרצה לדרוש כי בהסתברות של $1-\delta$ נקבל כי $e_S(h)$ ו – $e_P(h)$ הינם קרובים עד כדי $penalty$, ולכן נרצה חסם על ε.

כמו כן, נשתמש בנתון כי $\delta = 0.1$, וע"י מציאת חסם על ε מתוך האי-שיוויון האחרון נקבל –

$$4\ \left(\frac{en}{2k}\right)^{2k} e^{\frac{-n\varepsilon^2}{8}} \le \delta = 0.1 \qquad \Leftrightarrow \cdots \Leftrightarrow\ \ \varepsilon \ge \sqrt{\frac{8}{n}\cdot\left(2k\cdot ln\left(\frac{en}{k}\right) + ln\left(\frac{4}{0.1}\right)\right)}$$

ולכן נגדיר את פונק' ה – $penalty$ כך – $\boxed{penalty(h) = \sqrt{\frac{8}{n}\cdot\left(2k\cdot ln\left(\frac{en}{k}\right) + ln\left(\frac{4}{0.1}\right)\right)}}$.

הגרפים המתקבלים –



עבור הטעות האמיתית$(true\ error)$ - ה – k הטוב ביותר מתקבל עבור $k = 3$, וכן זהו **אותו ה – k האופטימלי** מהסעיף הקודם.
עבור הטעות האמפירית$(empiric\ error)$ - ה – k הטוב ביותר מתקבל עבור $k^* = 10$, וזהו זהו **אותו ה – k^*** מהסעיף הקודם, שאינו אופטימלי בהכרח.
עבור ה – $penalty$, קיבלנו כי ה – k הטוב ביותר מתקבל עבור $k = 1$, עבורו מתקבל $penalty$ מינימלי. פונק' ה- $penalty$ הינה פונק' מונוטונית עולה לפי הפרמטר k, ולכן $penalty$ מינימלי מתקבל עבור k קטן ככל הניתן.
לפי עיקרון ה-SRM, נבחר את ה- k שמביא למינימום את הסכום של הטעות האמפירית וה $penalty$–, ולכן **במקרה זה נקבל כי ה-k האופטימלי הינו $k = 3$, וזה כאמור אינו $k^* = 10$ שנבחר בסעיף הקודם ע"י ERM**, ע"י בחירה של k כך שהטעות האמפירית תהיה מינימלית. לפיכך, כיוון שראינו בסעיף ב' כי ההיפותזה האופטימלית מכילה אכן $k = 3$ קטעים, נסיק כי עיקרון ה-SRM מספק תוצאה דיי אופטימלית, ובפרט תוצאה טובה יותר מאשר קיבלנו בסעיף ד' ע"י שימוש באלגוריתם ה-ERM כפי שהוא.

סעיף ו.

לפי שיטת ה – $holdout - validation$, הפלט הינו –

```
RESULT OF HOLD-OUT:  3
ERROR OF  0.14333333333333334
Intervals are:
[ 0.0006650462038847449 , 0.19903942038289962 ]
[ 0.4001604403010063 , 0.5984344986363617 ]
[ 0.800475106452517 , 0.9995344858855774 ]
```

קיבלנו כי ה-k הטוב ביותר מתקבל עבור $k = 3$, עם שגיאה אמפירית של – 0.143, **והקטעים שהתקבלו הינם מאוד קרובים להיפותזה האופטימלית שחישבנו בסעיף ב'** – $[0.8, 1]$, $[0.4, 0.6]$, $[0,0.2]$. ניתן לראות כי שיטת ה – $holdout$ עבדה בצורה דיי מיטבית, והפלט ממנה הינה היפותזה קרובה להיפותזה האופטימלית שחישבנו, ולכן נסיק כי שיטת ה – $holdout$ הינה שיטה מצוינת.
התוצאות שקיבלנו בסעיף זה אכן מתכתבות עם המסקנה התיאורטית שהוכחנו בסעיף ב' – ההיפותזה האופטימלית מכילה 3 קטעים והם $[0.8, 1]$, $[0.4, 0.6]$, $[0,0.2]$, וכאמור התוצאות שהתקבלו הן קרובות מאוד לכך.