מבוא ללמידה חישובית – תרגיל בית 2

שם : דורי רימון

ת"ז : 323996843

<u>שאלות תיאורטיות:</u>

שאלה 1:

בהינתן מדגם אימון (train data) שסימנו S, האלגוריתם יפעל באופן הבא:

* אם קיים $z \in S$ עם לייבל 1, נחזיר את הפונק' h_z
* אחרת, נחזיר את הפונק' h^-.

קל לראות כי זהו אכן אלגוריתם ERM שכן השגיאה האמפירית שלו היא תמיד 0, ובפרט מינימלית (מינימום).

אכן, במידה וקיים $z \in S$ עם לייבל 1, הוא יחיד (מהנחת realizability), ומכאן שהפונק' h_z תתן שם שגיאה אמפירית 0.

במידה ולא קיים z כזה, אז h^- תתן גם שגיאה אמפירית 0.

אזי, בכל מקרה השגיאה האמפירית היא תמיד 0.

שאלה 2:

צ"ל: מחלקה H היא PAC learnable $\Leftrightarrow$ H היא PAC learnable in expectation

$\Leftarrow$ נניח כי מחלקה H היא PAC learnable. כלומר ק"מ אלגוריתם A כך שלכל $\varepsilon, \delta > 0$, ק"מ $N(\varepsilon, \delta)$ כך שלכל $n \geq N(\varepsilon, \delta)$ ולכל התפלג' P מתקיים $\mathbb{P}[e_P(A(S_n)) \leq \varepsilon] \geq 1 - \delta$.

לכל $n \geq N(\varepsilon, \delta)$ מתקיים אם כן,

מנוסחת התוחלת השלמה

$$\mathbb{E}[e_P(A(S_n))] = \underbrace{\mathbb{E}[e_P(A(S_n)) \mid e_P(A(S_n)) \leq \varepsilon]}_{\leq \varepsilon} \cdot \underbrace{\mathbb{P}[e_P(A(S_n)) \leq \varepsilon]}_{\leq 1} + \underbrace{\mathbb{E}[e_P(A(S_n)) \mid e_P(A(S_n)) > \varepsilon]}_{\leq 1} \cdot \underbrace{\mathbb{P}[e_P(A(S_n)) > \varepsilon]}_{\leq \delta} \leq$$

מההנחה

$$\leq \varepsilon \cdot 1 + 1 \cdot \delta = \varepsilon + \delta$$

אם כן, לכל $a \in (0,1)$, נבחר $\bar{\varepsilon} + \bar{\delta} = a$. עבור $\bar{\varepsilon}, \bar{\delta}$ אלו, מתקיים אם כן כי עבור $n \geq N(\bar{\varepsilon}, \bar{\delta})$,

$$\mathbb{E}[e_P(A(S_n))] \leq a$$

וקיבלנו את הדרוש. האלגוריתם A הנתון לנו מעיד כי H היא PAC learnable in expectation.

⇒ נניח כי H היא PAC learnable in expectation עם אלגוריתם A

ופונק' $N : (0,1) \to \mathbb{N}$ מתאים.

כעת קודם נראה כי A הנ"ל מעיד כי H היא PAC learnable.

ואם כן עלינו למצוא $\bar{N} : (0,1) \times (0,1) \to \mathbb{N}$.

יהיו $\varepsilon, \delta \in (0,1)$. נראה כי יתקיים $\mathbb{P}[e_P(A(S_n)) > \varepsilon] \le \delta$.

נשים לב כי

$$\mathbb{P}[e_P(A(S_n)) > \varepsilon] \le \frac{\mathbb{E}[e_P(A(S_n))]}{\varepsilon} \le \frac{a}{\varepsilon}$$

מההנחה .. אי-שוויון מרקוב

אי-שוויון זה נכון עבור כל $a \in (0,1)$ ו- $n \ge N(a)$.

אזי אם ניקח $a = \varepsilon \cdot \delta$, מתקיים כי עבור $\bar{N}(\varepsilon, \delta) = N(\varepsilon\delta)$

יתקיים הדרוש.

וקיבלנו את הרצוי.

שאלה 3:

נוכיח כי $VC\dim(H_k) = 2k$.

* ראשית, עלינו להראות כי קיימת קבוצה בגודל $2k$ שאותה ניתן לנתץ באמצעות H_k. למעשה, אין חשיבות בבחירת הנקודות כאן, ולכן תהי S_{2k} קבוצה כלשהי של $2k$ נקודות מהקטע $[0,1]$.

מכיוון שלכל היותר k אינטרוולים. תהי $y_1, \ldots, y_{2k}$ תיוג כלשהו של $S_{2k} := \{x_1, \ldots, x_{2k}\}$ לקלאסיפיים. B מעל לתיוג כולל בהכרח לפחות כי ניתן להציג כל תיוג כזה בעזרת היפותזה מהמחלקה.

אכן, לכל קבוצה "רציפה", כלומר העתקות אחת ואחרי השניה של היותר k, של נקודות עם תיוג 1, נקצה קטע (אינטרוול).

ישנן לכל היותר k קבוצות כאלו, ולכן הדבר אפשרי.

* כעת נרצה להראות כי כל קבוצה בגודל $2k+1$ לא ניתן לנתץ עם H_k. תהי קבוצה כלשהי $S_{2k+1} = \{x_1, \ldots, x_{2k+1}\}$ (נניח בה"כ כי הקבוצה ממוינת).

נתבונן בתיוגים $\{1, 0, 1, 0, \ldots, 1\}$, כלומר אינדקסים אי-זוגיים מתויגים ב-1 (הספירה ואינדקסים מתחילה מ-1).

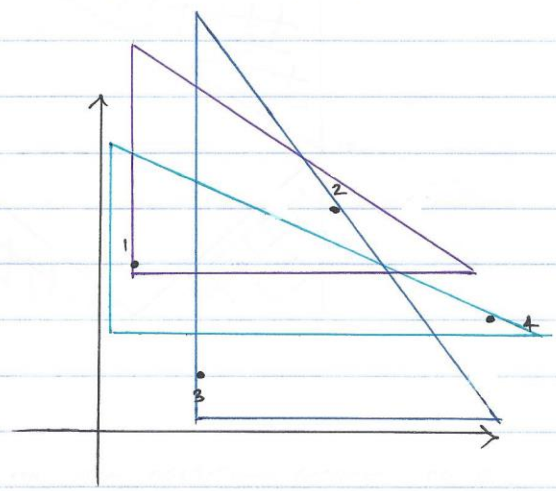
לא קיימת היפותזה ב-H_k אשר נותנת תיוג זה. ואכן, h מ-B אינטרוול להתחיל לפני נקודה אי-זוגית ולהסתיים לפני הנקודה הזוגית הבאה. לכן יש צורך ב-$k+1$ אינטרוולים.

לכן הוכחנו כי $VC\dim(H_k) = 2k$ כפי שרצינו.

שאלה 4:

נראה כי $vC\ dim(H) = 4$.

* ראשית, נתבונן בקבוצת 4 הנק' הבאה:



נחלק למקרים:

- אם נקודה אחת קיבלה את התיוג 1, פשוט נכיל אותה במשולש קטן.
- אם שתי נקודות קיבלו את התיוג 1, קל לראות טוב כי ניתן להקיף במשולש כנדרש מבלי לכלול את הנקודה הרביעית.
- עבור השאר, ניתן לראות כי 6 עם כל טוב להכיל במשולש אחד. זכרו את הנק' הנותרות. לצורך העניין ציירתי 3 משולשים, להדגמה עבור 3 המקרים הנותרים.

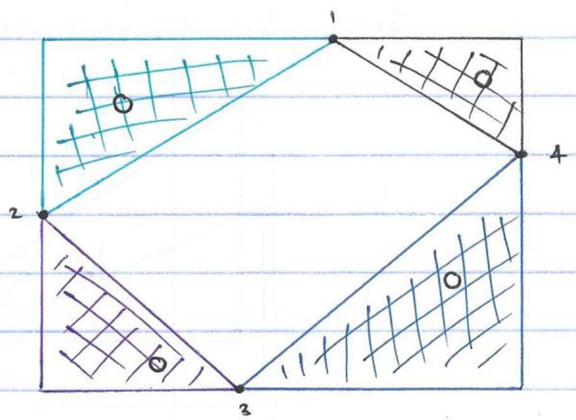
$(1,4) = \bullet$, $(1,2) = \bullet$, $(2,3) = \bullet$

* כעת נראה כי בסט של 5 נקודות, שלא ניתן לנתץ. לצורך כך נתבונן בסט כלשהו של 5 נקודות, ונסמן ב- $\tilde{x}$ את הנק' בעלת הגדולה ושל הקטנה ביותר בכל אחד מהצירים.

ראשית, אם הנקודות מסודרות בצורה כזו, שבה כאשר מחברים ביניהן בקו מקיף הצורה יוצרת קמורה, אז בהכרח לא ניתן לנתץ, שכן אם הנקודה הפנימית היא עם תיוג 0 וכל החיצוניות 1, לא ניתן ליצור הפרדה

מתאימות עם מעגלה.

אם כן, הנקודה $\tilde{x}$ ממוקמת באחד הבאים:



כאשר בתוך אחד המשולשים המסומנים, כמו אחת מהנקודות המסומנות כך 0.

עבור $\tilde{x}$ הממוקמת באזור הכחול / סגול / כחול, שם ניתן לראות כי ישנן רק את תיוג 0 ושל הצבע 1. אכן, לא תהיה במעגל מקבילה לצירים את 4 הנקודות הנבחרות תורה את הקצה $\tilde{x}$.

עבור $\tilde{x}$ בתוך השחור, נבחן את המקרה בו $\tilde{x}$ עם תיוג 1, והנקודה הבאה בתור משמאלה עם תיוג 0.
עבור כך ישנן 2 אפשרויות:
1) הקו המקביל לציר y מפריד ביניהן - במקרה זה נקודה 2 תמיד מחוץ למעגל.
2) הציר מפריד ביניהן. אם יש 2 מתוך שנקודה 4 תקבל תיוג 0, הקו המקביל לציר x חייב להפריד בין $\tilde{x}$ ונקודה 4, ובמקרה זה, נקודה 3 תמיד תהיה מחוץ למעגל.

אם כן, הוכחנו כי כל קבוצה בת 5 נקודות לא ניתנת לניתוץ.

שאלה 5:

אז יהיו δ, ε. נשים לב כי

$$\mathbb{P}\left[\exists h \in H,\ |e_P(h) - e_S(h)| > \sqrt{\frac{1}{2|S|} \ln \frac{2k \cdot |H_i|}{\delta}}\ \right] \le \frac{\delta}{k}$$

$$\Rightarrow \mathbb{P}\left[\exists h \in H,\ 1 \le i \le k\ ,\ |e_P(h) - e_S(h)| > \text{—"—}\ \right] \le$$

$$\underbrace{\le}_{\text{חסם האיחוד}} \sum_{i=1}^{k} \underbrace{\frac{\delta}{k}}_{\text{משני - השוויון}} = k \cdot \frac{\delta}{k} = \delta$$

ואז המאורע המשלים מתקיים בהסתברות $\ge 1-\delta$ כרצוי. המוקף הבא.

אם כן, יש להוכיח את אי השוויון המוקף הבא.

הוכחנו בהרצאה כי

$$\mathbb{P}\left[\exists h \in \tilde{H},\ |e_P(h) - e_S(h)| > \varepsilon\right] = \mathbb{P}\left[\sup_{h \in \tilde{H}} |e_P(h) - e_S(h)| > \varepsilon\right] \le$$

$$\le 2 \cdot |\tilde{H}| \cdot e^{-\underset{n=|S|}{n}\varepsilon^2}$$

אז נותר להוכיח כי

$$\circledast\ 2 \cdot |H_i| \cdot \exp\left(-2 \cdot \underbrace{|S|}_{n} \cdot \left(\underbrace{\sqrt{\frac{1}{2|S|} \cdot \ln \frac{2k \cdot |H_i|}{\delta}}}_{\varepsilon \text{ במקרה שלנו}}\right)^2\right) = \frac{\delta}{k}$$

$$\exp\left(-\cancel{2|S|} \cdot \frac{1}{\cancel{2|S|}} \cdot \ln \frac{2k \cdot |H_i|}{\delta}\right) = \frac{\delta}{2k \cdot |H_i|}$$

$$\Rightarrow \circledast = \frac{\delta}{2k \cdot |H_i|} \cdot 2 \cdot |H_i| = \frac{\delta}{k}$$

וקיבלנו את הרצוי.

ק) הפעם נניח כי האינדקס המוחזר ע- SRM הוא $\tilde{i}$. נסמן

$$e_P(SRM(S)) - e_P(h^*) \le e_S(ERM_{\tilde{i}}(S)) + \sqrt{\frac{1}{2n} \cdot \ln \frac{2k \cdot |H_{\tilde{i}}|}{\delta}} - e_P(h^*)$$

נזכר כי SRM
מחזיר ERM
בהסתברות לפחות
ההנחה עם 1-δ , מהסעיף הקודם
הוא בוחר

$$\le e_S(ERM_{i^*}(S)) + \sqrt{\frac{1}{2n} \cdot \ln \frac{2k \cdot |H_{i^*}|}{\delta}} - e_P(h^*) \le$$

SRM ממזער
את שני המחוברים
הנוגעים, כלל...

$$\le e_S(h^*) + \sqrt{\frac{1}{n} \ln \frac{2k \cdot |H_{i^*}|}{\delta}} - e_P(h^*) \le$$

ERM, כמו כן הוא,
מביא למינימום את
השגיאה האמפירית, כלל.

$$\le 2\sqrt{\frac{1}{n} \ln \frac{2k \cdot |H_{i^*}|}{\delta}} \le \varepsilon$$

מהסעיף הקודם,
בהסתברות 1-δ

נרצה לבחור n מספיק גדול.

$$\Rightarrow 2\ln(2k \cdot |H_{i^*}| / \delta) \le n\varepsilon^2$$

$$\Rightarrow n \ge \frac{2}{\varepsilon^2} \ln \frac{2k \cdot |H_{i^*}|}{\delta}$$

וקיבלנו את התנאי.

מטלה מעשית:

א.

א) ראשית נשים לב כי $e_P(h) = \mathbb{E}_P[\Delta_{zo}(h(x), y)] = L(h)$

ההפסד כפי שהוגדר בהרצאה 1.

כמו כן, אנו במקרה הבינארי, ונתונה לנו ההתפלגות האמיתית P. לכן, ניתן להשתמש ב- MAP, ואז כפי שהוכחנו בהרצאה 1, מהווה את ההיפותזה האופטימלית, כלומר ה $\arg\min_h L(h)$.

כפי שהוכחנו, ההיפותזה האופטימלית היא $h(x) = \arg\max_{y\in\{0,1\}} \mathbb{P}(Y=y \mid X=x)$

ובמקרה הספציפי הזה:

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

ההיפותזה האופטימלית

$$\mathbb{P}(y=1 \mid x) = 0.8 > 0.2 = \mathbb{P}(y=0 \mid x) \Leftarrow x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]$$

$$\Rightarrow h(x) = 1$$

$$\mathbb{P}(y=1 \mid x) = 0.1 < 0.9 = \mathbb{P}(y=0 \mid x) \Leftarrow x \in (0.2, 0.4) \cup (0.6, 0.8)$$

$$\Rightarrow h(x) = 0$$

ב.

$$e_p(h_I) = \mathbb{E}_p(\Delta_{zo}(h(X), Y)) =$$ ב2

$$= \sum_{y \in \{0,1\}} \int_0^1 \mathbb{P}(X = x, Y = y) \cdot \Delta_{zo}(h(x), y) \cdot dx =$$

$$= \int_0^1 \mathbb{P}(X = x) \cdot \mathbb{P}(Y = 1 \mid X = x) \cdot \Delta_{zo}(h(x), 1) +$$

$$+ \int_0^1 \mathbb{P}(X = x) \cdot \mathbb{P}(Y = 0 \mid X = x) \cdot \Delta_{zo}(h(x), 0)$$

נתמקד בחישוב של הביטוי הראשון, אך זה תקף עבור השניה, נסמן $y = y$:

$$\int_0^1 \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y \mid X = x) \cdot \Delta_{zo}(h(x), y) =$$

$$= \sum_{l_i, u_i \in I} \int_{l_i}^{u_i} \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y \mid X = x) \cdot \Delta_{zo}(1, y) +$$

$$I = \{[l_1, u_1], [l_2, u_2], \ldots, [l_k, u_k]\}$$

$$+ \sum_{u_i, l_{i+1} \in I} \int_{u_i}^{l_{i+1}} \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y \mid X = x) \cdot \Delta_{zo}(0, y) =$$

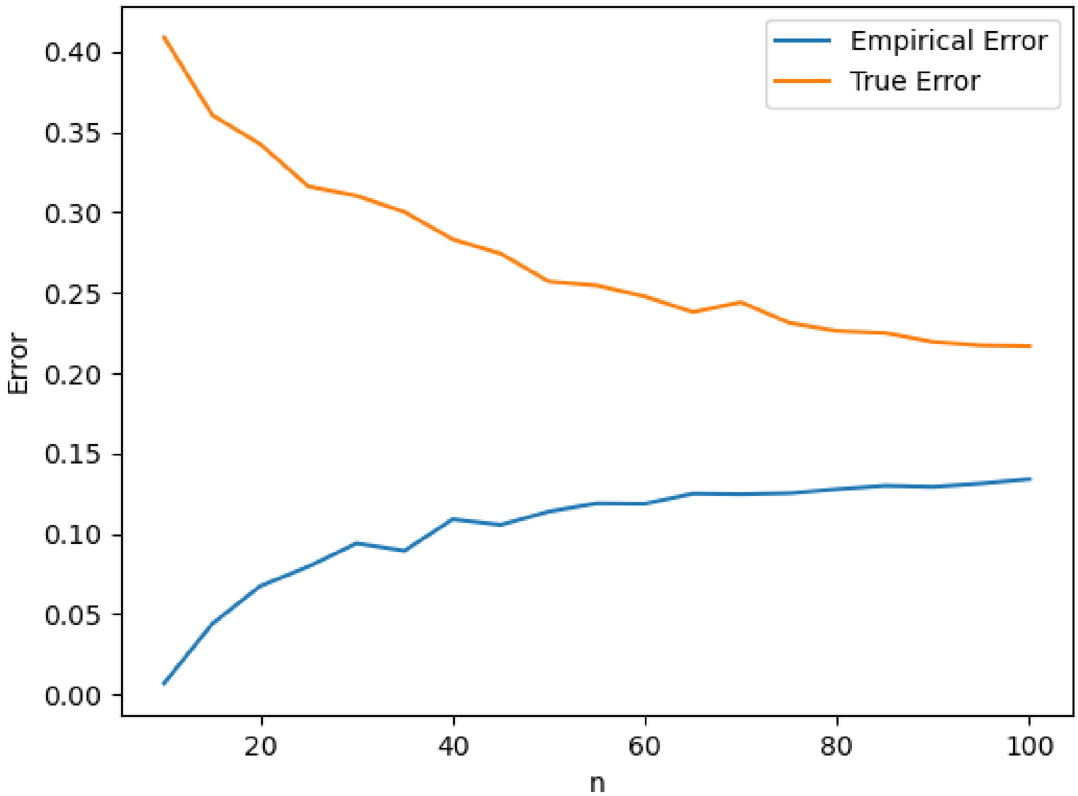
בין כל סיום של קטע

לתחילת הקטע הבא

$$= \sum_{l_i, u_i \in I} (u_i - l_i) \cdot \int_{l_i}^{u_i} \mathbb{P}(Y = y \mid X = x) \cdot \Delta_{zo}(1, y) +$$

$$+ \sum_{u_i, l_{i+1} \in I} (l_{i+1} - u_i) \cdot \int_{u_i}^{l_{i+1}} \mathbb{P}(Y = y \mid X = x) \cdot \Delta_{zo}(0, y)$$

מציאת פתרון אחד, חושבה הפונק' true_error המתארת את $e_p(h_I)$.

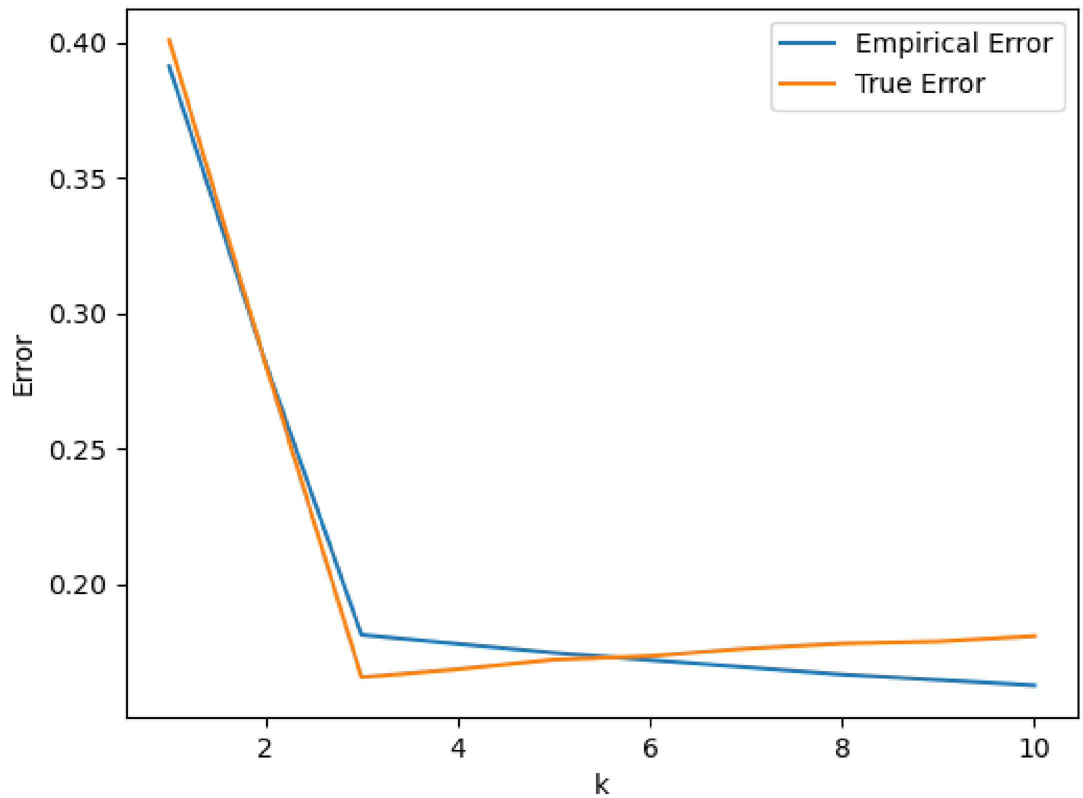


ניתן לראות כי באופן סכמתי, השגיאה האמיתית יורדת ב n. הסיבה לכך היא כי ככל ש n עולה, החישוב האמפירי של התוחלת (שהוא החישוב המתבצע כשמחשבים את השגיאה האמיתית) מתקרב להסתברות האמיתית P לכך שנקודה תוגרל מחוץ לקטע. הסיבה לשיפוע השלילי המוזכר, היא כי תחילה לאלגוריתם ERM אין מספיק נתונים על מנת לראות כיצד מתנהגת ההתפלגות P. מכאן שהשגיאה ביחס ל P תהיה גדולה יחסית. אך ככל ש n עולה, יתקרב האלגוריתם להתפלגות האמיתית. במקרה ה realizable, אף יגיע לשגיאת האמת בסופו של דבר.

בנוגע לשגיאה האמפירית, ככל ש n עולה, יש יותר נקודות שעל האלגוריתם לסווג, בעודו מוגבל ב $k = 3$ אינטרוולים עמם הוא יכול לעדן את החלוקה. מכאן שאט אט השגיאה תלך ותעשה גדולה יותר.

דבר אחרון לשים לב אליו, הוא כי אנו במקרה מעניין, המזכיר מקרה realizable. אמנם אין פונקציית איחוד אינטרוולים שיצרה את ה data, אך הפונקציה שסיווגה את הנקודות קרובה מאוד לכזו. העבודה היא לפי מקטעים, ובהסתברות גבוהה מאוד (0.8) נקודה בתוך מקטע מסווגת ל 1, ובהסתברות גבוהה מאוד (0.9) נקודה מחוץ לקטע מסווגת ל 0. על כן על אף ה "רעש" שנוסף, ניתן לראות כי השגיאה האמפירית מתקרבת לשגיאה האמיתית, כפי שהיינו מצפים במקרה ה realizable.

ג.



בנוגע לגרף השגיאה האמיתית, ניתן לראות כי הוא מקבל מינימום ב $k = 3$. עובדה זו מסתדרת עם הניתוח התיאורטי שעשינו בסעיף א.

נשים לב להתנהגות גרף השגיאה האמפירית. ניגע בנושא זה בשני מישורים :

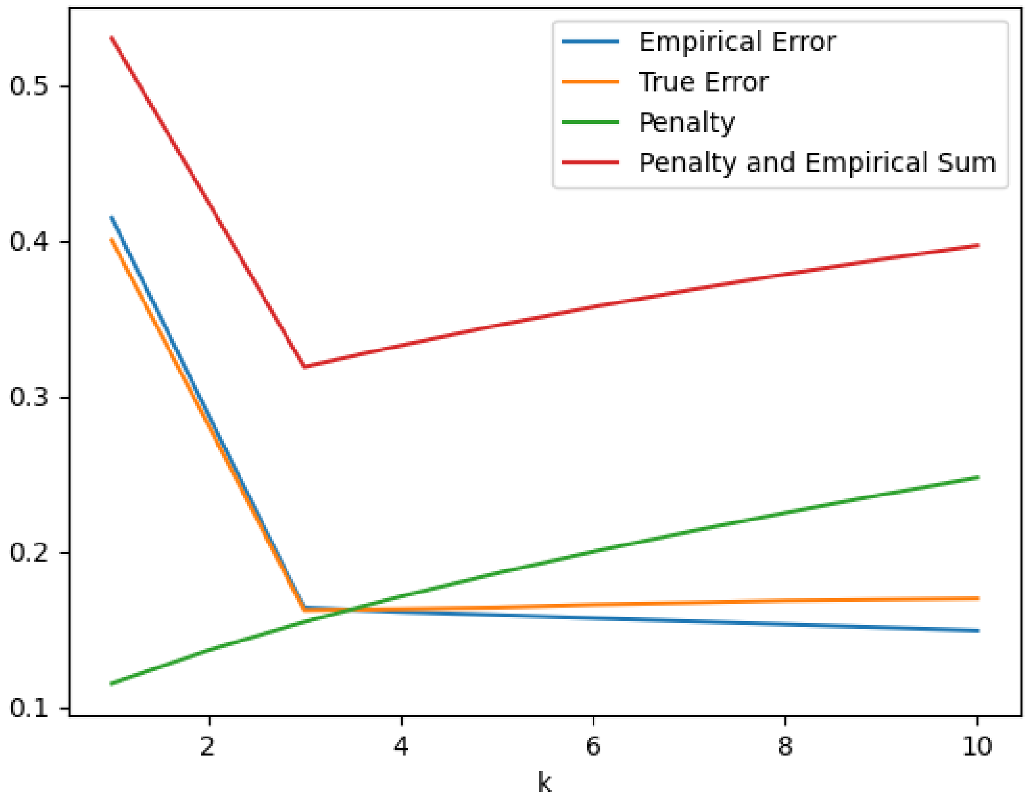
ראשית, הגרף יורד. כמובן כי עובדה זו צפויה, שכן ככל שנגדיל את k, כך נוכל לעדן את החלוקה לאינטרוולים על ה train data שלנו, ולהוריד את השגיאה האמפירית. מדובר במדע מדויק – קל לראות כי השגיאה של אלגוריתם ERM כפונקציה של k היא מונוטונית יורדת.

שנית, ואולי אף מעניין יותר, היא העובדה כי שיפוע הגרף קבוע יחסית ביחס ל k השונים, אך משתנה באופן דרסטי בנקודה $k = 3$. במילים אחרות, התרומה השולית של k כאשר $k > 3$ קטנה מהותית מזו שעד הנקודה הזו. נשים לב כי $k = 3$ היא בדיוק נקודת המינימום של גרף השגיאה האמיתית.

נזכר כי ראינו בכיתה חסם על השגיאה האמיתית, כסכום של penalty ושל השגיאה האמפירית. בסעיף הבא אעמיק בכך קצת יותר, אך ניתן כבר כאן לחזות בכך שכאשר $k > 3$ נקבל overfitting.

בנוגע ל k^*, גרף השגיאה האמפירית יורד, ובפרט נקודת המינימום שלו היא קצה הקטע, כלומר $k^* = 10$. עם זאת, זה אינו k אופטימלי, כפי שגרף השגיאה האמיתית מרמז.

ד.

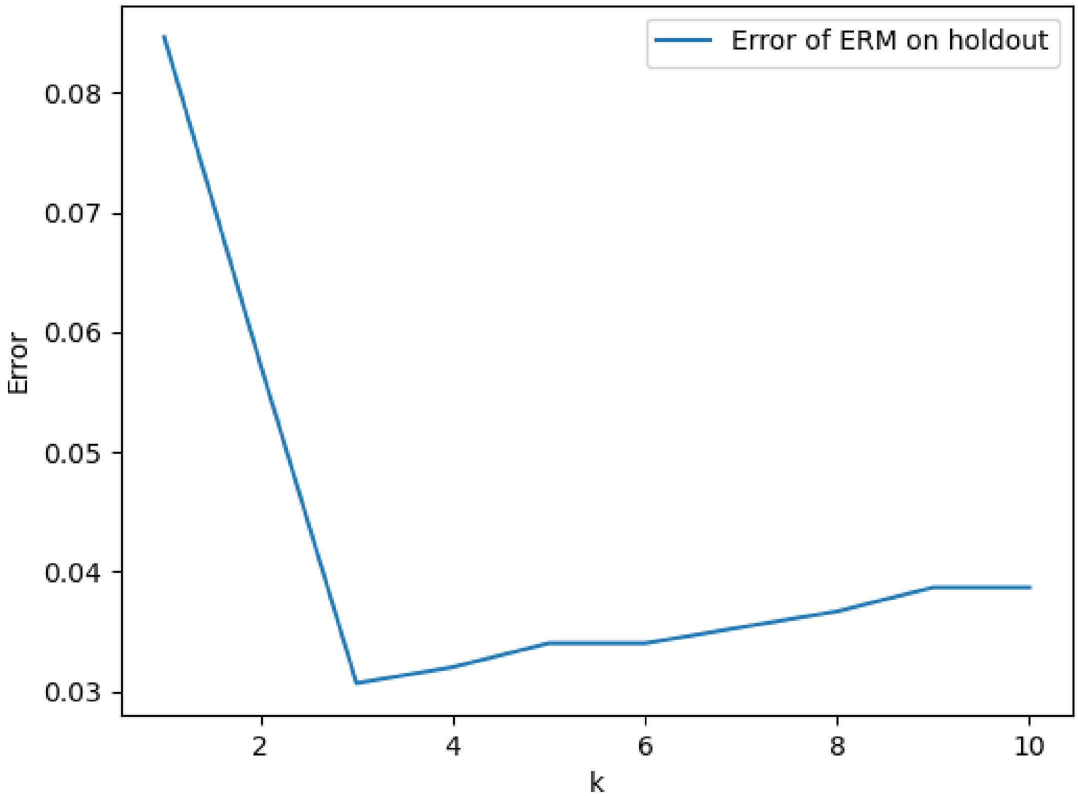


כעת ניתן לראות את התמונה השלמה, עליה התחלנו לדון בסעיף הקודם. נקודת המינימום של גרף הסכום (שהוא החסם העליון שלנו על השגיאה האמיתית) היא $k^* = 3$. כמובן כי זהו k עדיף על זה שהתקבל בסעיף הקודם.

התחלנו לדון בסעיף הקודם על שינוי השיפוע של גרף השגיאה האמפירית, וכאן ניתן לראות ממש, שכיוון שהתרומה השולית של הגדלת k כאשר $k > 3$ כה קטנה, היא לא "משתלמת" לעומת ה penalty שעלינו לשלם, וכך מתקבל overfitting.

על מנת להשלים את התמונה, עבור $k < 3$ נשים לב כי אמנם ה penalty שאנו משלמים נמוך, אך זו לא נקודה אופטימלית כיוון שהתרומה השולית של k לגרף השגיאה האמפירית כה גדולה. מכאן שבחלק זה נקבל underfitting.

ה.



במקרה זה המינימום מתקבל עבור $k = 3$. כפי שראינו, H_3 היא אכן המשפחה ממנה מגיע האופטימום.