<u>הערה לבודק!!!</u> כתבתי את התרגיל כולו ב One Note , אך הייתה לי בעיה עם הייצוא לPDF. הפתרון הזמני שחשבתי עליו היה לעשות סקרינשוטים ולהדביק בוורד. מצטער על אי הנוחות.

שב 2

Friday, 11 November 2022 10:45

איל גרינברג 207129792

Theory Questions

2. **(15 points) PAC in Expectation**. Consider learning in the realizable case. We say a hypothesis class $\mathcal{H}$ is ***PAC learnable in expectation*** using algorithm A if there exists a function $N(a) : (0,1) \to \mathbb{N}$ such that $\forall a \in (0,1)$ and for any distribution P (realizable by $\mathcal{H}$), given a sample set S such that $|S| \geq N(a)$, it holds that,

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

Show that $\mathcal{H}$ is PAC learnable *if and only if* $\mathcal{H}$ is PAC learnable in expectation (Hint: For one direction, use the law of total expectation. For the other direction, use Markov's inequality).

צ"ל: $\mathcal{H}$ למידה PAC בתוחלת $\Leftrightarrow$ $\mathcal{H}$ למידה PAC.

<u>$\Leftarrow$</u>: נניח ש-$\mathcal{H}$ למידה PAC בתוחלת ע"י אלגוריתם A כלשהו.
כלומר לכל $a\in(0,1)$ קיים $N(a)$ כך שלכל התפלגות P ריאליזבילית ע"י $\mathcal{H}$, בהינתן מדגם גודל S שנדגם מההתפלגות P והגודל שלו $N(a)\leq|S|$ מתקיים: $\mathbb{E}\left[e_P(A(S))\right]\leq a$.

נרצה להראות ש-$\mathcal{H}$ למידה PAC ע"י אותו אלג' A. כלומר, לכל $\varepsilon,\delta\in(0,1)$ קיים $N(\varepsilon,\delta)$ כך שלכל התפלגות P ריאליזבילית ע"י $\mathcal{H}$, בהינתן מדגם S כזה שנדגם מ-P והוא בגודל $n\geq N(\varepsilon,\delta)$ מתקיים: $Pr\left(e_P(A(S))>\varepsilon\right)\leq\delta$.

ראשית, מהגדרת ה-test error: $e_P(A(S))=\mathbb{E}\left[\Delta_{0-1}(A_S(X),Y)\right]$ נשים לב שמדובר בתוחלת של מ"מ שמקבל את ערכים 0 או 1 ולכן $e_P(A(S))$ הוא מס' אי שלילי, לכן ניתן להפעיל את אי-שוויון מרקוב עליו.

נקבל שלכל $a\in(0,1)$ ולכל גודל מדגם שגדול מ-$N(a)$ מתקיים:

$$Pr\left[e_P(A(S))>\varepsilon\right]\overset{\text{Markov}}{\leq}\frac{\mathbb{E}\left[e_P(A(S))\right]}{\varepsilon}\overset{\mathcal{H}\text{ למידה PAC בתוחלת}}{\leq}\frac{a}{\varepsilon}$$

לכן בפרט לכל $a\in(0,1)$ ובפרט עבור $a=\varepsilon\cdot\delta$ לכל $\varepsilon,\delta\in(0,1)$
(מכיוון שמכפלת מספרים בין 0 ל-1 גם היא בין 0 ל-1). כלומר נקבל $Pr\left[e_P(A(S))>\varepsilon\right]\leq\frac{a}{\varepsilon}=\delta$ כנדרש.

לכן עבור $a=\delta\cdot\varepsilon$, $N(\varepsilon,\delta)=N(a)=N(\varepsilon\cdot\delta)$ נקבל ש-$\mathcal{H}$ למידה PAC ע"י אותו אלג' A.

$\Rightarrow$:

נשים לב שאם ניקח בזוג המאורעות הזרים והמשלימים $\{e_p(A(S)) > \varepsilon\}$, $\{e_p(A(S)) \le \varepsilon\}$.
לפי משפט ההסתברות השלמה נקבל לכל $\delta, \varepsilon \in (0,1)$ עבור $n > N(\varepsilon,\delta)$ נקבל:

$$E[e_p(A(S))] = \underbrace{E[e_p(A(S)) \mid e_p(A(S)) \le \varepsilon]}_{*} \cdot \underbrace{Pr[e_p(A(S)) \le \varepsilon]}_{*} +$$
$$\underbrace{E[e_p(A(S)) \mid e_p(A(S)) > \varepsilon]}_{*} \cdot \underbrace{Pr[e_p(A(S)) > \varepsilon]}_{*}$$

נרצה לחסום את התוחלת, הנה לכן נחסום כל אחד מהגורמים המרכיבים אותה:

$* \le \delta$ מכיוון ש $\mathcal{H}$ למידה PAC.

$* \le 1$ מכיוון שבכל שגיאה בכיוון השני $e_p(A(S))$ היא תוחלת על מ"מ שמקבל ערכים 0 או 1 ולכן תוחלת המ"מ הזה חסומה מלמעלה ע"י 1.

$* \le 1$ כי מדובר בהסתברות.

$* \le \varepsilon$ כי בהינתן ש $e_p(A(S)) \le \varepsilon$ מובן שתוחלתו לא גדולה מ-ε.

סה"כ נקבל: $E[e_p(A(S))] \le \varepsilon \cdot 1 + 1 \cdot \delta = \varepsilon + \delta$

כדי לסיים את ההוכחה נראה שלכל $a \in (0,1)$ ולכן לכל $n \ge N(a)$ יתקיים: $E[e_p(A(S))] \le a$.
ואכן: לכל $a \in (0,1)$ ניתן לבחור כמו $a = \varepsilon' + \delta'$ עבור $\varepsilon', \delta' \in (0,1)$ כלשהם (לפי ההוכחה של התוחלת הקודמת)
לכן אם ניקח $n \ge N(a) = N(\varepsilon' + \delta') = N(\varepsilon', \delta')$ נקבל $E[e_p(A(S))] \le a$ כנדרש. ■

3. **(15 points) Union Of Intervals.** Determine the VC-dimension of $\mathcal{H}_k$ - the subsets of the real line formed by the union of k intervals (see the programming assignment for a formal definition of $\mathcal{H}$). Prove your answer.

$VCdim(\mathcal{H}_k) = 2k$

נוכיח $VCdim(\mathcal{H}_k) \geq 2k$:

עלינו למצוא תת קבוצה $C \subseteq \mathcal{X}$ בגודל 2k המנותצת ע״י $\mathcal{H}_k$.

תהי קבוצה כלשהי: $C=\{x_1,x_2,\dots,x_{2k} \mid x_1<x_2<\dots<x_{2k}\}$

יהיו $s_1,s_2,\dots,s_{2k} \in \{0,1\}$ דיכוטומיה כלשהי.

נגדיר אינטרוול לכל רצף של תיוגים 1 במובן הבא: נאמר רצף $s_i,s_{i+1},\dots,s_j$ כך שכל התיוגים הם 1 נגדיר אינטרוול שיתחיל ב-x_i ויסתיים ב-x_j. מכאן שכל תיוגים חיוביים באורך 1 יכוסה ע״י אינטרוול שמכיל נקודה בודדת.

מכיוון שמס׳ הרצפים של תיוגים 1 המקסימלי בקבוצה בגודל 2k הוא k,

למשל: $s_1\ s_2\ s_3\ s_4 \dots s_{2k-3}\ s_{2k-2}\ s_{2k-1}\ s_{2k}$ זו סדרה הגרועה של דיכוטומיות,

$$\underbrace{0\ \underline{1}\ 0\ \underline{1}\ \dots\ 0\ \underline{1}\ 0\ \underline{1}}_{\text{סה״כ k צ׳ים}}$$

נקבל שכל רצף אפשרי של כל דיכוטומיה ניתן לכסות בעזרת k אינטרוולים. כנדרש.

נוכיח $VCdim(\mathcal{H}_k) \leq 2k$:

עלינו להראות שכל קבוצה $C \subseteq \mathcal{X}$ בגודל 2k+1 לא ניתן לנתץ ע״י $\mathcal{H}_k$.

אכן לכל קבוצה $C=\{x_1,\dots,x_{2k+1} \mid x_1<x_2<\dots<x_{2k+1}\}$ לא קיים ב-$\mathcal{H}_k$ h המייצג את הדיכוטומיה שמגדירה את התיוגים

$s_1\ s_2\ s_3\ s_4 \dots s_{2k-3}\ s_{2k-2}\ s_{2k-1}\ s_{2k}\ s_{2k+1}$

$1\ 0\ 1\ 0 \quad 1\ 0\ 1\ 0\ 1$

זאת משום שיש בקבוצה זו k+1 לייבלים שהם 1 ובין כל שני לייבלים 1 יש לייבל 0, כלומר נדרש לכסות k+1 אינטרוולים שונים לכסות את k+1 הנקודות $x_1,x_3,x_5,\dots,x_{2k-1},x_{2k+1}$. בעוד שכל היפותזה כאמור יש רק k אינטרוולים להשתמש בהם. ■

4. **(15 points) Prediction by polynomials.** Given a polynomial $P : \mathbb{R} \to \mathbb{R}$ define the hypothesis $h_P : \mathbb{R}^2 \to \{0, 1\}$ by,

$$h_P(x_1, x_2) = \begin{cases} 1 & P(x_1) \geq x_2 \\ 0 & \textit{otherwise.} \end{cases}$$

Determine the VC-dimension of $\mathcal{H}_{poly} = \{h_P \mid P \text{ is a polynomial}\}$. You can use the fact that given n distinct values $x_1, ..., x_n \in \mathbb{R}$ and $z_1, ..., z_n \in \mathbb{R}$ there exists a polynomial P of degree $n-1$ such that $P(x_i) = z_i$ for every $1 \leq i \leq n$.

$VCdim(\mathcal{H}_{poly}) = \infty$.

על מנת להוכיח שמימד ה-VC של מחלקת $\mathcal{H}_{poly}$ אינסופי, נוכיח שלכל n טבעי יש תת-קבוצה של $\mathcal{X}$ בגודל n שמנותצת ע"י $\mathcal{H}_{poly}$. כלומר: $\forall n \in \mathbb{N}.\ VCdim(\mathcal{H}_{poly}) \geq n$. ומכך ינבע שמימד הVC של $\mathcal{H}_{poly}$ אינסופי.

נגדיר תת-קבוצה של $\mathcal{X}$: $C = \{(x_1, z_1), (x_2, z_2), \dots, (x_n, z_n) \mid x_1, \dots, x_n \in \mathbb{R}$ שונים זה מזה$\}$

ומתקיים כמובן $|C| = n$.

קל לראות שניתן לקבל את הדיכוטומיה $[1, 1, \dots, 1]$ בפשטות ע"י הפולינום הבא, עבור כל הנקודות.

כלל קיים פולינום P ממעלה $n-1$ כך ש: $\forall 1 \leq i \leq n.\ P(x_i) = z_i$.

ולכן מתקיים עבור הפולינום P: $\forall i: h_P(x_i, z_i) = 1$.

כעת, נראה שכל תת-סדרה של סדרת תיוגים באורך n שכוללת לפחות רק תיוגים 0 ו-1 שונים בדיכוטומיה ששונה מ-$\mathcal{H}_{poly}$. עם לעשות נוכל שניתן לקבל כל דיכוטומיה אפשרית, כלומר $\mathcal{H}_{poly}$ מנתצת את C.

לכל סדרת אינדקסים $1 \leq i_1 < i_2 < \dots < i_k \leq n$ נוכל שקיים המסווג h_P ב-$\mathcal{H}_{poly}$ כך שלכל אינדקס i_j בסדרה זו מתקיים $h_P(x_{i_j}, z_{i_j}) = 0$ ועבור כל אינדקס שלא בסדרה זו h_P יחזיר תיוג 1.

תהי סדרת אינדקסים $1 \leq i_1 < i_2 < \dots < i_k \leq n$ שכזו, אז נחסיר $\varepsilon > 0$ מכל z_{i_j} $(1 \leq j \leq k)$ בקבוצה C, ונקבל את הקבוצה החדשה C_ε:

$$C_\varepsilon = \{(x_{i_1}, z_{i_1} - \varepsilon), (x_{i_2}, z_{i_2} - \varepsilon), \dots, (x_{i_k}, z_{i_k} - \varepsilon)\} \cup \{(x_i, z_i) \in C \mid i \notin \{i_1, \dots, i_k\}\}$$

C_ε היא קבוצה של מספרים ב-$\mathbb{R}^2$ בגודל n כאשר הקואורדינטות הראשונות בכל אחת זהות לקואורדינטות הראשונות מכל אחת בקבוצה C. לכן ניתן להפעיל את הטענה על איברי C_ε ולקבל שקיים פולינום P_ε כך ש:

$$\forall i \notin \{i_1, \dots, i_k\}.\ P_\varepsilon(x_i) = z_i \Rightarrow \forall i \notin \{i_1, \dots, i_k\}.\ h_{P_\varepsilon}(x_i, z_i) = 1$$

$$\forall i \in \{i_1, \dots, i_k\}.\ P_\varepsilon(x_i) = z_i - \varepsilon < z_i \Rightarrow \forall i \in \{i_1, \dots, i_k\}.\ h_{P_\varepsilon}(x_i, z_i) = 0$$

לכן $\mathcal{H}_{poly}$ מנתצת כל תת-קבוצה בגודל n של $\mathcal{X}$ ולכן $VCdim(\mathcal{H}_{poly}) = \infty$. ■

1. **(15 points) PAC learnability of ℓ_2-balls around the origin.** Given a real number $R \geq 0$ define the hypothesis $h_R : \mathbb{R}^d \to \{0,1\}$ by,

$$h_R(\mathbf{x}) = \begin{cases} 1 & \|\mathbf{x}\|_2 \leq R \\ 0 & \textit{otherwise.} \end{cases}$$

Consider the hypothesis class $\mathcal{H}_{ball} = \{h_R \mid R \geq 0\}$. Prove directly (without using the Fundamental Theorem of PAC Learning) that $\mathcal{H}_{ball}$ is PAC learnable in the realizable case (assume for simplicity that the marginal distribution of X is continuous). How does the sample complexity depend on the dimension d? Explain.

נציע אלגוריתם A_{ball} שלומד את $\mathcal{H}_{ball}$ במקרה הריאליזבילי.

נסמן ב h_{R_0} את ההיפותזה האמיתית, כלומר את ההיפותזה שעבורה מתקיים $Y = h_{R_0}(X)$.

האלגוריתם A_{ball}, בהינתן n נקודות ב $\mathbb{R}^d$ ימצא את הכדור B הקטן ביותר שקונסיסטנטי עם המדגם, כלומר מקיים עבורו $\ell_S(h_B) = 0$.

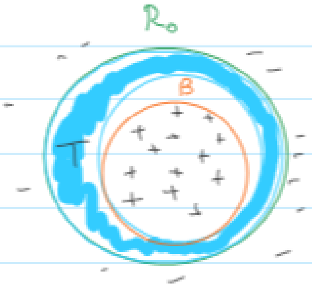
נשים לב ש h_B יכול לטעות רק ע"י סוג נקודות חיוביות כשליליות, אלה הן הנקודות אשר נמצאות בטבעת $R_0 \setminus B$. לכן שגיאת האמת של A_{ball} תהיה: $e_P(h_B) = \mathbb{E}[\Delta(h_B(X), Y)] = Pr[h_B(X) \neq Y] = Pr[R_0 \setminus B]$.

נניח שמתקיים $Pr[R_0] > \varepsilon$ כי אם $Pr[R_0] \leq \varepsilon$ נקבל: $e_P(h_B) = Pr[R_0 \setminus B] \leq Pr[R_0] \leq \varepsilon$ וסיימנו.

כעת נגדיר "כדור חלול" T כך שיהיה מוגבלת הכדור R_0 ו"יתחת" לכיוון פנים הכדור (ראשית הצירים) עד שהמסה של T תהיה $P[T] = \varepsilon$. נשים לב שהדבר "תקין" כי נתון שX מגיע מהתפלגות רציפה.

שימו לב להמחשה במקרה בדו מימדי כש B, R_0 הם מעגלים ו T הוא טבעת.

בשרטוט כנראה מאוד גרוע אבל מדובר בסה"כ במקרה דומה לתרגיל עם המלבנים מתרגול 2, אך הכיוון לרעיון דומה.



נראה כי אם קיימת נקודה $x^{(i)}\in T$ אזי בהכרח $\hat{g}$ לא הינו נק' חיובית, ושכן $T\subseteq R_0$. מכיוון שעני זו
תפ"ה מתאים B על P, כי הוא קונסיסטנטי. צריך להתקיים שהשפה בין הכדורים $R_0 B$ נמצאת בתוך הכדור החלל
T, כלומר $R_0\setminus B\subseteq T$.

לכן נקבל: $e_P(h_B)=Pr[R_0\setminus B]\le Pr[T]\le\varepsilon$

נסכם את שקיבלנו: $\underbrace{\exists 1\le i\le n.\ x^{(i)}\in T}_{A}\Rightarrow\underbrace{e_P(h_B)\le\varepsilon}_{B}$

כאשר שמתקיים $\neg B\Rightarrow\neg A$

ולכן: $e_P(h_B)>\varepsilon\Rightarrow\forall 1\le i\le n.\ x^{(i)}\notin T$

מכיוון שה LHS גורר את RHS מתקיים $Pr[LHS]\le Pr[RHS]$.

כלומר, $Pr[e_P(h_B)>\varepsilon]\le Pr[\forall 1\le i\le n.\ x^{(i)}\notin T]$

עכשיו נחשב את RHS:

$$Pr[x^{(i)}\notin T]=1-Pr[x^{(i)}\in T]=1-\varepsilon \qquad 1-x\le e^{-x}$$

$$Pr[\forall 1\le i\le n.\ x^{(i)}\notin T]\underset{iid}{=}\prod_{i=1}^{n}Pr[x^{(i)}\notin T]=\prod_{i=1}^{n}1-\varepsilon=(1-\varepsilon)^n\le(e^{-\varepsilon})^n=e^{-\varepsilon\cdot n}$$

נרצה למצוא את הגודל n עבורו $Pr[e_P(h_B)>\varepsilon]\le\delta$ כלומר n שיקיים:

$$e^{-\varepsilon n}\le\delta\iff-\varepsilon n\le\ln\delta\iff n\ge-\frac{1}{\varepsilon}\cdot\ln\delta\iff n\ge\frac{1}{\varepsilon}\ln\left(\frac{1}{\delta}\right)$$

לכן נקבע סיבוכיות דגימה $N(\varepsilon,\delta)=\frac{1}{\varepsilon}\ln(\frac{1}{\delta})$.

לסיכום, הוכחנו שקיים $N(\varepsilon,\delta)=\frac{1}{\varepsilon}\ln(\frac{1}{\delta})$ כך שלכל $\varepsilon,\delta\in(0,1)$ ולכל התפלגות P ריאליזבילית
על R^d עם סוג $h_{R_0}\in\mathcal{H}_{ball}$. כאשר נריץ את אלג' A_{ball} על $n\ge N(\varepsilon,\delta)$ דגימות מתוך שנדגמו iid מההתפלגות
P, הוא יחזיר היפותזה $h_B\in\mathcal{H}_{ball}$ כך ש: $Pr[e_P(h_B)>\varepsilon]\le\delta$

ולכן $\mathcal{H}_{ball}$ למידה PAC ע"י האלגוריתם A_{ball}

כמו כן, סיבוכיות הדגימה אינה תלויה במימד d.

הסיבה לכך היא שהמימד VC של $\mathcal{H}_{ball}$ הוא 1. נראה זאת:

$VCdim(\mathcal{H}_{ball})\ge1$: לכל $x\in R^d$ כך ש $\|x\|_2=r>0$ קיימים R_0,R_1 כך ש $R_0<r<R_1$.

אזי, $h_{R_0}(x)=0$, $h_{R_1}(x)=1$ הן שתי היפותזות ב $\mathcal{H}_{ball}$ המנתצות את הסינגלטון $\{x\}\subseteq\mathcal{X}$.

$VCdim(\mathcal{H}_{ball})\le1$: לכל שתי נקודות שונות $x_1,x_2\in R^d$ כך שבה"כ מתקיים $\|x_1\|_2\le\|x_2\|_2$.

לא קיימת ב $\mathcal{H}_{ball}$ היפותזה כך ש $h_R(x_1)=0\wedge h_R(x_2)=1$ כי על מנת ש $h_R(x_2)=1$
יתקיים, צריך ש $\|x_2\|_2\le R$ אבל $\|x_1\|_2\le\|x_2\|_2 \Leftarrow \|x_1\|_2\le R \Leftarrow h_R(x_1)=1$ בסתירה.

מכיוון ש $VCdim(\mathcal{H}_{ball})=1$ ולא תלוי בגודל d, כך גם הסיבוכיות דגימה לא תלויה בו.

Programming Assignment

(a) **(8 points)** Assume that the true distribution $P[x, y] = P[y|x] \cdot P[x]$ is as follows: x is distributed uniformly on the interval $[0, 1]$, and

$$P[y = 1|x] = \begin{cases} 0.8 & \text{if } x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] \\ 0.1 & \text{if } x \in (0.2, 0.4) \cup (0.6, 0.8) \end{cases}$$

and $P[y = 0|x] = 1 - P[y = 1|x]$. Since we know the true distribution P, we can calculate $e_P(h)$ precisely for any hypothesis $h \in \mathcal{H}_k$. What is the hypothesis in $\mathcal{H}_{10}$ with the smallest error (i.e., $\arg\min_{h\in\mathcal{H}_{10}} e_P(h)$)?

בהרצאה ראינו שהסיווג מינימלי עם סף הפסד 0-1 המסווג האופטימלי הוא מסווג MAP.

כלומר מסווג $h(x)$ מהצורה:

$$h(x) = \begin{cases} 1 & \text{if } Pr[Y=1|X=x] > Pr[Y=0|X=x] \\ 0 & \text{if } Pr[Y=1|X=x] \leq Pr[Y=0|X=x] \end{cases}$$

עפ״י ההתפלגות הנתונה של X,Y מתקיים:

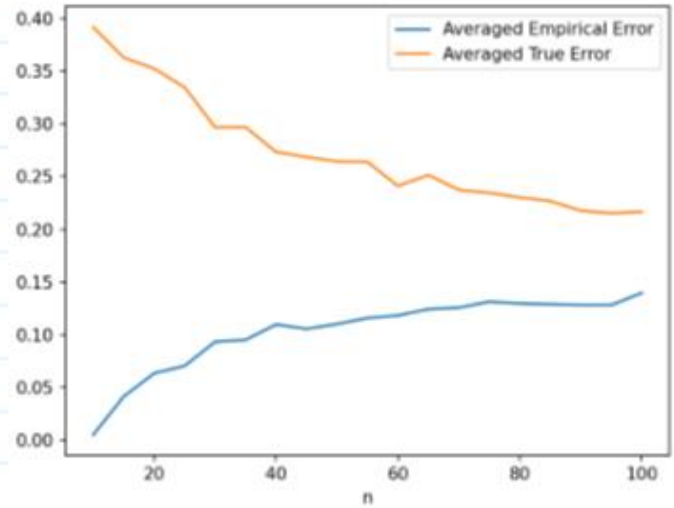
$$Pr[Y=1|X=x \in [0,0.2] \cup [0.4,0.6] \cup [0.8,1]] = 0.8$$
$$Pr[Y=1|X=x \in [0.2,0.4] \cup [0.6,0.8]] = 0.1$$
$$Pr[Y=0|X=x \in [0,0.2] \cup [0.4,0.6] \cup [0.8,1]] = 1-0.8 = 0.2$$
$$Pr[Y=0|X=x \in [0.2,0.4] \cup [0.6,0.8]] = 1-0.1 = 0.9$$

ולכן המסווג האופטימלי יהיה:

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

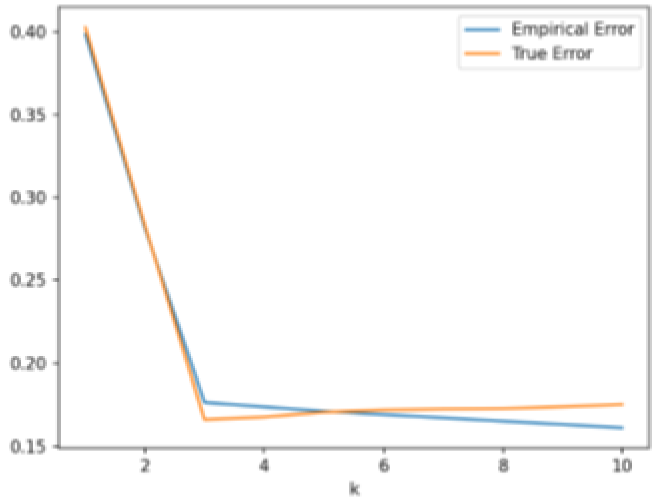
נשים לב שאין תלות כאן במס' האינטרוולים k של מחלקת ההיפותזות $\mathcal{H}_k$ לשגיאה, לכן $h^*(x)$ הוא המסווג האופטימלי לכל מחלקת היפותזות $\mathcal{H}_k$, ובפרט גם עבור מס' ההיפותזות $\mathcal{H}_{10}$.

(b) **(8 points)** Write a function that, given a list of intervals I, calculates the true error $e_P(h_I)$. Then, for $k = 3$, $n = 10, 15, 20, \ldots, 100$, perform the following experiment $T = 100$ times: (i) Draw a sample of size n and run the ERM algorithm on it; (ii) Calculate the empirical error for the returned hypothesis; (iii) Calculate the true error for the returned hypothesis. Plot the empirical and true errors, averaged across the T runs, as a function of n. Discuss the results. Do the empirical and true errors decrease or increase with n? Why?



- ניתן לראות שהשגיאה האמפירית עולה כשגודל המדגם גדל. הסיבה לכך היא שככל שיש יותר נקודות יותר סביר שנטעה, כלומר יתקבלו תיוג '0' באינטרוולים $[0,0.2] \cup [0.4,0.6] \cup [0.8,1]$ או תיוג '1' באינטרוולים המשלימים. כמו כן, כאשר n קטן מאוד, למשל $n=10$ השגיאה האמפירית יכולה להיות קטנה גם כן כי לשלושה אינטרוולים ניתן לקבל $e_S(h)=0$ כי זה לא קשה עבור המדגם הקטן הזה.
- ניתן לראות שהשגיאה האמיתית קטנה ככל ש-n גדל. הסיבה לכך היא שככל שגודל המדגם גדל כך נתקרב להתפלגות האמיתית P.
- כמו כן, ניתן לראות שתי השגיאות מתקרבות זו לזו, כלומר מאוד קרוב לחוק המספרים הגדולים, כאשר מגדילים את גודל המדגם הממוצע ישאף לתוחלת.

(c) **(8 points)** Draw a sample of size $n = 1500$. Find the best ERM hypothesis for $k = 1, 2, \ldots, 10$, and plot the empirical and true errors as a function of k. How does the error behave? Define k^* to be the k with the smallest empirical error for ERM. Does this mean the hypothesis with k^* intervals is a good choice?



ניתן לראות ששתי השגיאות יורדות כאשר $1 \leq k \leq 3$. לאחר מכן הטעות האמיתית מתחילה לעלות והטעות האמפירית ממשיכה לרדת ונקבל $k^* = 10$.

הסיבה לכך שה-k הטוב ביותר הוא 10 היא שמתקיים over fit, לאור הפרמטר בהן יש יותר אינטרוולים המודל יותר טוב למדגם וככה מקטין את הטעות האמפירית.

זה מצב לא רצוי כמובן כי גורם ל variance ב tradeoff bias-variance גדל במצב של over fitting ובכך מגדיל את השגיאה באמת. בדומה, הכללה $k > 3$ אינטרוולים החזירה חלוקה פחות טובה מהחלוקה האופטימלית ל-$k=3$ בגלל כך שאנו גוזרים k. במקום אחת, ה-"sweet spot" המתקבל כאשר $k=3$ שם הטעות האמיתית הכי מינימלית.

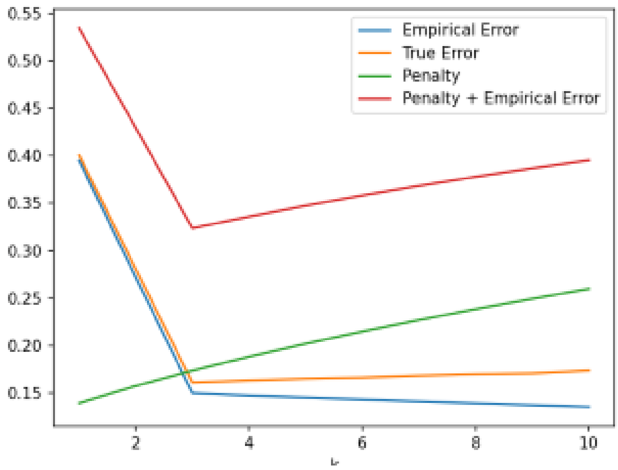
(d) **(8 points)** Now we will use the principle of structural risk minimization (SRM), to search for a k that gives a good test error. Let $\delta = 0.1$:

- Use to following penalty function:

$$2\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_k) + \ln\frac{2}{\delta}}{n}}$$

- Draw a data set of $n = 1500$ samples, run the experiment in (c) again, but now plot two additional lines as a function of k: 1) the penalty for the best ERM hypothesis and 2) the sum of penalty and empirical error.
- What is the best value for k in each case? is it better than the one you chose in (c)?

ראשית, אציין כי הבנתי מהניסוח בשאלה שעלי להוסיף משקולות $w_1,\dots,w_{10} \geq 0$, $\sum_{k=1}^{10} w_k = 1$
על מנת לשמור על החסם בנלמן של $\delta = 1-0.9$ עבור כל 10 מחלקות ההיפותזות, ולא רק עבור
כל מחלקת היפותזות בנפרד. מכיוון שמס' המחלקות הנו סופי (10 מח'), אפשר למשקל כל δ באופן
אחיד, כלומר לקבוע $w_k = 0.1$, $1 \leq k \leq 10$.
במצב ולא היה צריך להוסיף משקולות (כלומר ה-$\delta = 0.1$ כבר מגלם את ה-penalty הרצוי), אז כל
δ ב-penalty היה שווה 0.1 ולא $w_k \cdot 0.1 = 0.1 \cdot 0.1 = 0.01$.
כלומר מדובר בסך הכל בקנס כך שמבחינת ההסבר מהגרף אין הבדל.



כפי שניתן לראות שה-penalty עולה ב-k וזה הגיוני לפי מה שלמדנו בהרצאה, ככל שמחלקת
ההיפותזות יותר מורכבת (VCdim גדול יותר) כך ה-penalty גדל (יש צורך בקנס גדול יותר עבור מחלקה יותר
מורכבת). בנוסף, ניתן לראות שהגרף של השגיאה האמפירית + העונש מתנהג בצורה דומה לשגיאה
האמיתית, כשהוא למעשה חוסם אותה מלמעלה (בהסתברות $1-\delta$) כפי שלמדנו.
ה-k האופטימלי עבור השגיאה האמפירית + העונש הוא $k=3$, כפי שראינו לאורך כל הסעיפים ההיפותזה
האופטימלית שייכת למחלקה של $k=3$ כלומר זהה לא $k^*=10$ מהסעיף הקודם.

(e) **(8 points)** Here we will use holdout-validation to search for a $k \in \{1, ..., 10\}$ that gives good test error. Draw a data set of $n = 1500$ samples and use 20% for a holdout-validation. Choose the best hypothesis and discuss how close this gets you to finding the hypothesis with optimal true error.

```
The intervals of the best hypothesis are:  [(0.0012140697609384182, 0.2007047337147445), (0.3998050485581141, 0.5985120185325733), (0.7990006373824257, 0.9999530506112453)]

The best k is:  3
```

כצפוי, ההיפותזה הטובה ביותר התקבלה עבור k=3. ניתן לראות ששלושת האינטרוולים מאוד קרובים לאינטרוולים בהיפותזה האופטימלית מסעיף א': $[[0,0.2],[0.4,0.6],[0.8,1]]$