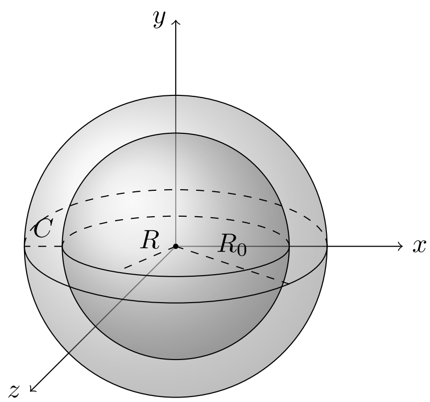
Intro to Machine Learning: Assignment #2

May 2, 2023

Dor Bourshan
11111

Theory Questions

1 PAC learnability of ℓ_2-balls around the origin



הגדרה: בהינתן מספר ממשי $R \geq 0$ מגדירים את ההיפותזה $h_R : \mathbb{R}^d \to \{0,1\}$ כ :

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

נסמן :

$$\mathcal{H}_{ball} = \{h_R \mid R \geq 0\}$$

נוכיח באופן ישיר כי מחלקת ההיפותזות $\mathcal{H}_{ball}$ למידת PAC במקרה ה realizable .

אנו צריכים להראות שקיים אלגוריתם $\mathcal{A}$ כך ש לכל $\varepsilon, \delta > 0$ קיימת פונקציה $N : (0,1) \times (0,1) \to \mathbb{N}$ כך שלכל $n \geq N(\varepsilon, \delta)$ מתקיים

$$\forall P \in R(\mathcal{H}_{ball}) \quad : \quad \Pr\left[e_P\left(\mathcal{A}\left(S_n\right)\right) \leq \varepsilon\right] \geq 1 - \delta$$

או לחלופין

$$\Pr\left[e_P\left(\mathcal{A}\left(S_n\right)\right) > \varepsilon\right] < \delta$$

נגדיר את האלגוריתם הבא $\mathcal{A}_{ball}$:

- בהינתן n דגימות S_n נגדיר את הקבוצה S_1 להיות קבוצת הדגימות מתוייגות $'1'$
- נגדיר

$$R = \max_{\mathbf{x} \in S_1} \|\mathbf{x}\|_2$$

- נחזיר $h_R(x)$

מהיותה של המחלקה realizable השגיאה היחידה היא עבור $\mathbf{x}$ המתוייגים $'1'$ כך ש $R < \|\mathbf{x}\|_2 \leq R_0$

נגדיר

$$R_0 \backslash R = \{\mathbf{x} \mid R < \|\mathbf{x}\|_2 \leq R_0\}$$

אנו מעוניינים לתת הערכה הסתברותית לשגיאה של האלגוריתם, כלומר לתת הערכה הסתברותית לנקודות שנופלות ב $R_0 \backslash R$ כיוון שנקודות אלו מתוייגות כ $'0'$ למרות שאמורות להיות מתוייגות כ $'1'$. נשים לב כי $R \leq R_0$ שוב, מהיותה של המחלקה realizable

אנו מניחים כי

$$\Pr[R_0] > \varepsilon$$

אחרת

$$\Pr[R_0 \backslash R] \leq \Pr[R_0] \leq \varepsilon$$

כלומר

$$\Pr[e_P(\mathcal{A}(S_n)) \leq \varepsilon] = 1$$

נגדיר נקודה $C \leq R$ כך ש

$$\Pr[R_0 \backslash C] = \varepsilon \tag{1}$$

אם קיימת נקודה $\mathbf{x}$ בדגימה, המתוייגת כ $'1'$ כך ש $\|\mathbf{x}\|_2 \geq C$, אז מעצם הגדרת $\mathcal{A}_{ball}$ מתקיים כי

$$R_0 \backslash R \subseteq R_0 \backslash C$$

ולכן

$$\Pr[R_0 \backslash R] \leq \Pr[R_0 \backslash C] = \varepsilon$$

(#)נזכור כי עבור דגימה S_n אף נקודה לא נפלה ב $R_0 \backslash R = \{\mathbf{x} \mid R < \|\mathbf{x}\|_2 \leq R_0\}$, נקודות המתוייגות כ $'0'$ אינן יכולות ליפול בתחום זה כיוון שהמחלקה realizable וכן מעצם הגדרת האלגוריתם $R = \max_{\mathbf{x} \in S_1} \|\mathbf{x}\|_2$

אחרת כל הנקודות המתוייגות כ $'1'$ בדגימה, מקיימות $\|\mathbf{x}\|_2 < C$, כלומר

$$\Pr[R_0 \backslash R] > \varepsilon$$

אם כן :

$$\Pr[e_P(h_R) > \varepsilon] \overset{(\#)}{=} \Pr[\forall i \;\; \|\mathbf{x}\|_2 \notin (C, R]] \overset{\text{i.i.d}}{=} \prod_{i=1}^{n} \Pr[\|\mathbf{x}\|_2 \notin (C, R]] \overset{1}{=} (1-\varepsilon)^n \leq e^{-\varepsilon n} \leq \delta$$

נפתור עבור n :

$$e^{-\varepsilon n} \leq \delta \iff -\varepsilon n \leq \ln(\delta) \iff n \geq \frac{1}{\varepsilon} \ln\left(\frac{1}{\delta}\right)$$

נשים לב כי החסם איננו תלוי ב d. נוכל להסביר זאת בשתי דרכים (שיכולות להיות שקולות) :

- מתקיים כי $VC\dim(\mathcal{H}_{ball}) = 1$ (כי בהינתן שתי נקודות, אם הן בעלות אותו נורמה, מהגדרת המחלקה הן לא ניתנות לניפוץ כלומר הן יכולות לקבל את אותו label בלבד. אם הן בעלות נורמה שונה, אז לא נוכל להגדיר את הנקודה בעלת הנורמה הגדולה כמתוייגת $'1'$ ואת הנקודה השניה כ $'0'$ מעצם הגדרת המחלקה. באופן ברור נקודה אחת ניתנת לניפוץ)
- בלקיחת נורמה הורדנו את מימד הבעיה ל 1.

$\square$

2 PAC in Expectation

נניח בשאלה כי אנו מדברים על המקרה ה realizable .

הגדרה : נאמר ש מחלקת היפותזות $\mathcal{H}$ היא $PAC\ learnable\ in\ expectation$ באמצעות אלגוריתם $\mathcal{A}$ אם קיימת פונקציה $N(a): (0,1)^2 \to \mathbb{N}$ כך ש $\forall a \in (0,1)$ ולכל התפלגות $P \in R(\mathcal{H})$, בהינתן קבוצת דגימה S כך ש $|S| \geq N(a)$ מתקיים כי

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

נוכיח כי $\mathcal{H}$ היא PAC learnable אם ורק אם $\mathcal{H}$ היא PAC learnable in expectation

הוכחה : נוכיח בשני כיוונים :

- $\Longleftarrow$ $\mathcal{H}$ היא PAC learnable .

יהיי $a \in (0,1)$ ויהיו $\varepsilon, \delta > 0$ כך ש $a = \varepsilon + \delta$. מתקיים כי לכל $n \geq N(\varepsilon, \delta)$ ולכל $P \in R(\mathcal{H})$ כי :

$$\Pr\left[e_P\left(\mathcal{A}(S_n)\right) \leq \varepsilon\right] \geq 1 - \delta$$

ובאופן שקול

$$\Pr\left[e_P\left(\mathcal{A}(S_n)\right) > \varepsilon\right] < \delta$$

נסמן ב $\mathcal{H}_{bad}^{\varepsilon}(P)$ את קבוצת ההיפותזות המקיימות :

$$\mathcal{H}_{bad}^{\varepsilon}(P) = \{h \in \mathcal{H} \mid e_P(h) > \varepsilon\}$$

מהגדרת הקבוצה מתקיים כי

$$\Pr\left[e_P\left(\mathcal{A}(S_n)\right) > \varepsilon\right] = \Pr\left[\mathcal{A}(S_n) \in \mathcal{H}_{bad}^{\varepsilon}\right] < \delta \quad (2)$$

$$\Pr\left[e_P\left(\mathcal{A}(S_n)\right) \leq \varepsilon\right] = \Pr\left[\mathcal{A}(S_n) \notin \mathcal{H}_{bad}^{\varepsilon}\right] \geq 1 - \delta \quad (3)$$

מנוסחאת התוחלת השלמה נקבל כי

$$\begin{aligned}
&\mathbb{E}\left[e_P\left(\mathcal{A}(S)\right)\right] = \mathbb{E}\left[\mathbb{E}\left[e_p(h) \mid \mathcal{A}(S) = h\right]\right] = \\
&\underbrace{\mathbb{E}\left[e_p(h) \mid \mathcal{A}(S) \in \mathcal{H}_{bad}^{\varepsilon}\right]}_{\leq 1} \cdot \underbrace{\Pr\left[\mathcal{A}(s) \in \mathcal{H}_{bad}^{\varepsilon}\right]}_{\text{From2}<\delta} + \underbrace{\mathbb{E}\left[e_p(h) \mid \mathcal{A}(S) \notin \mathcal{H}_{bad}^{\varepsilon}\right]}_{\text{From2}\leq\varepsilon} \cdot \underbrace{\Pr\left[\mathcal{A}(s) \notin \mathcal{H}_{bad}^{\varepsilon}\right]}_{\leq 1} \\
&\leq \delta + \varepsilon = a
\end{aligned}$$

- $\Longrightarrow$ $\mathcal{H}$ היא PAC learnable in expectation

יהיו $\varepsilon, \delta > 0$. יהיי $a \in (0,1)$ כך ש $a = \varepsilon \cdot \delta$. לכל $n \geq N(a)$ (פונקציה של ε, δ) ולכל התפלגות $P \in R(\mathcal{H})$ מתקיים

$$\Pr\left[e_P\left(\mathcal{A}(S)\right) > \varepsilon\right] \overset{\text{Chebyshev}}{\leq} \frac{\mathbb{E}\left[e_P\left(\mathcal{A}(S)\right)\right]}{\varepsilon} \leq \frac{a}{\varepsilon} = \frac{\varepsilon \cdot \delta}{\varepsilon} = \delta$$

□

3 Union Of Intervals

נגדיר את מרחב ההתפלגות המשותפת באופן הבא: $\mathcal{X} = [0,1]$ ו $\mathcal{Y} = \{0,1\}$

נגדיר קבוצת אינטרוולים: $I = \{[l_1,u_1],\ldots,[l_k,u_k]\}$: איחוד של k אינטרוולים כך שמתקיים:

$0 \le l_1 \le u_1 \le \ldots \le u_k \le 1$. לכל קבוצת אינטרוולים שכזו נגדיר היפותזה:

$$h_I(x) = \begin{cases} 1 & \text{if } x \in \bigcup_{i=1}^{k} [l_i,u_i] \\ 0 & otherwise \end{cases}$$

לבסוף נגדיר את מחלקת ההיפותזות הבאה:

$$\mathcal{H}_k = \{h_I \mid I = \{[l_1,u_1],\ldots,[l_k,u_k]\}, 0 \le l_1 \le u_1 \le \ldots \le u_k \le 1\}$$

נראה כי $VC\dim(\mathcal{H}_k) = 2k$

אנו נדרשים להראות שני שלבים:

1. תניתן לנפץ קבוצה בת $2k$ איברים.
2. לא ניתן לנפץ קבוצה גדולה יותר.

אם כן:

1. אם כן נראה כי יש קבוצה בת $2k$ איברים הניתנת לניפוץ. כלומר אנו נראה כי קיימת קבוצת נקודות (בגודל $2k$) שניתן לתייגן ב 2^{2k} דרכים שונות.

   יהיו $2k$ נקודות בקטע $[0,1]$ המקיימות

$$0 < x_1 < x_2 < \ldots < x_{2k} < 1$$

   תהיי $\mathcal{L}_1$ קבוצת הנקודות אותן אנו מעונויינים לתייג כ $'1'$. ובאופן זהה את $\mathcal{L}_0$ את קבוצת הנקודות שאנו מעוניינים לתייג כ $'0'$

   - אם $q = |\mathcal{L}_1| \le k$:
     מצפיפות הממשיים ניתן לעטוף כל נקודה $x_{i_j} \in \mathcal{L}_1$ בקטע $[l_{i_j},u_{i_j}]$ כך שלכל $1 \le r \le 2k$, כך ש $x_r \ne x_{i_j}$ מתקיים כי: $x_r \notin [l_{i_j},u_{i_j}]$
     עבור $k-q$ שנותר לקבוע את זהותם, מצפיפות הרציונלים קיים קטע $[l_*,u_*]$כך שלכל $1 \le i \le 2k$, כך ש מתקיים כי: $x_i \notin [l_*,u_*]$. נגדיר את כל הקטעים הנותרים כקטע זה.
     אם כך לכל $x_j \notin \mathcal{L}_1$ מתקיים כי תיוגו הוא $'0'$ וכן לכל $x_i \in \mathcal{L}_1$ מתקיים כי תיוגו $'1'$, כנדרש.
   - אם $q = |\mathcal{L}_1| > k$:
     נתייחס לנקודות סמוכות, קרי $x_{i_j}, x_{i_j+1}, \ldots, x_{i_j+r} \in \mathcal{L}_1$ כנקודה יחידה y_i, ועבור $x_j \in \mathcal{L}_1$ כך ש $x_{j-1}, x_{j+1} \notin \mathcal{L}_1$ נתייחס גם כן כ y_j.
     נסמן ב Y המוגדרת על ידי נקודות אלה. נטען כי $|Y| \le k$.

נניח בשלילה כי $|Y| \geq k+1$ כלומר אנו טוענים כי ישנן $k+1$ נקודות שאינן בסמיכות, כלומר לכל $y_i \in Y$ קיימת נקודה (אחת לפחות) כך ש $x_j \in \mathcal{L}_0$ המפרידה בין 2 נקודות ב Y. כדי לייצר הפרדה נדרשות לפחות k נקודות בקבוצה $\mathcal{L}_0$ אולם מתקיים כי $|\mathcal{L}_0| < k$. סתירה.

לפיכך $|Y| \leq k$ ואם כן חזרנו למקרה שכבר טופל, מצפיפות הממשיים נוכל לעטוף נקודות סמוכות ב $\mathcal{L}_1$ בקטע יחיד שלא מכיל אף נקודה אחרת מלבדן.

עברנו על כל האפשרויות, ונוכחנו לדעת שניתן לתייג את $2k$ הנקודות ב 2^{2k} דרכים שונות. אם כך הצלחנו לנתץ קבוצה המונה $2k$ איברים.

2. נראה כי לא ניתן לנתץ מדגם המכיל $r > 2k$ נקודות. נעיר כי מהגדרת הפונקציות במחלקת ההיפותזות, עבור נקודות בדגימה המקיימות $x_i = x_j$ עבור $i \neq j$, מתקיים כי התייוג של נקודות אלה זהה, ולכן לא ניתן לנתץ מדגם שכזה. אם כן נניח כי כל הנקודות במדגם שונות.

נראה שלא ניתן לתייג את הנקודות האי זוגיות במדגם ב $'1'$ ונסיים. כדי לתייג את הנקודות האי זוגיות אנו נדרשים לאינטרוול I_1 שיעטוף את x_1 ולא את x_2 ואינטרוול I_2 שיעטוף את x_3 ולא את $x_{2,}x_4$ וכו'

סה"כ נדרשים לפחות $k+1$ אינטרוולים שונים, למשל עבור מדגם בגודל $2k+1$

$$0 \leq l_1 \leq x_1 \leq u_1 < x_2 < l_2 \leq x_3 \leq u_2 < \ldots < l_{k+1} \leq x_{2k+1} \leq u_{k+1} \leq 1$$

אולם מחלקת ההיפותזות מורכבת מהיפותזות של k אינטרוולים. כלומר קבוצץ דגימה המקיימת שגודלה גדול מ $2k$ אינה ניתנת לניפוץ.

כנדרש.

□

4 Prediction by polynomials

בהינתן פולינום $P:\mathbb{R}\to\mathbb{R}$ נגדיר את ההיפותזה $h_P:\mathbb{R}^2\to\{0,1\}$ ע"י :

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

נגדיר את המחלקת ההיפותזות הבאה :

$$\mathcal{H}_{poly}=\{h_P\mid P\text{ polynomial a is}\}$$

נראה כי $VC\dim\left(\mathcal{H}_{poly}\right)=\infty$.

נשתמש בעובדה כי בהינתן $x_1,\ldots,x_n\in\mathbb{R}$ שונים ו $z_1,\ldots,z_n\in\mathbb{R}$ קיים פולינום P כך ש $\deg\left(P\right)=n-1$ כך ש לכל $1\leq i\leq n$ מתקיים כי $P\left(x_i\right)=z_i$

יהיי n.

נראה כי קיימת קבוצה בת n איברים הניתנת לניפוץ. נסמן

$$S=\left\{\begin{pmatrix}x_1^1\\x_2^1\end{pmatrix},\ldots,\begin{pmatrix}x_1^n\\x_2^n\end{pmatrix}\right\}$$

כך שמתקיים לכל $i\neq j$ כי $x_1^i\neq x_1^j$.

יהיו $y_1,\ldots,y_n$ התיוגים שאנו מעוניינים לייחס ל $\begin{pmatrix}x_1^1\\x_2^1\end{pmatrix},\ldots,\begin{pmatrix}x_1^n\\x_2^n\end{pmatrix}$ בהתאמה.

נראה שקיים פולינום P בדרגה $n-1$ כך ש h_P מתייג כנדרש. נגדיר פונקציה $\mathcal{L}:\{0,1\}\times\mathbb{R}^2\to\mathbb{R}$ באופן הבא :

$$\mathcal{L}\left(y_i,\left(x_1^i,x_2^i\right)\right)=\begin{cases}x_2^i & \text{if } y_i=1\\ x_2^i-1 & \text{if } y_i=0\end{cases}$$

קיים פולינום P כך ש $\deg\left(P\right)=n-1$ כך ש לכל $1\leq i\leq n$ מתקיים כי $P\left(x_1^i\right)=\mathcal{L}\left(y_i,\left(x_1^i,x_2^i\right)\right)$

נתבונן ב h_P :

$$h_P\left(x_1^i,x_2^i\right)=\begin{cases}1 & P\left(x_1^i\right)\geq x_2^i\\ 0 & otherwise\end{cases}=\begin{cases}1 & \mathcal{L}\left(y_i,\left(x_1^i,x_2^i\right)\right)\geq x_2^i\\ 0 & otherwise\end{cases}=\begin{cases}1 & y_i=1\\ 0 & otherwise\end{cases}$$

כלומר h_P מתייגת כנדרש.

הנ"ל נכון לכל n ולכן מתקיים כי

$$VC\dim\left(\mathcal{H}_{poly}\right)=\infty$$

Programing Assignment

1. Union Of Intervals

(a)

נניח כי ההתפלגות האמיתית $\Pr[x,y]=\Pr[y\mid x]\cdot\Pr[x]$ נתונה ע"י : $X\sim U([0,1])$

$$\Pr[Y=1\mid X=x]=\begin{cases}0.8 & \text{if } x\in[0,0.2]\cup[0.4,0.6]\cup[0.8,1]\\ 0.1 & \text{if } x\in(0.2,0.4)\cup(0.6,0.8)\end{cases}$$

מכך שאנו יודעים את ההתפלגות האמיתית אנו יודעים לחשב את $e_P(h)$ קרי השגיאה תחת h, לכל היפותזה $h\in\mathcal{H}_k$.

ההיפותזה בעלת השגיאה הקטנה ביותר במחלקה $\mathcal{H}_{10}$ קרי

$$h=\arg\min_{h\in\mathcal{H}_{10}} e_P(h)$$

היא

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

זאת תוצאה ישירה מהתרגיל הקודם בו הוכחנו כי

$$h=\arg\min_{f:\mathcal{X}\to\mathcal{Y}}\mathbb{E}\left[\ell_{0-1}(Y,f(X))\right]=e_P(h)$$

נתונה ע"י

$$h(x)=\arg\max_{i\in\mathcal{Y}}\Pr[Y=i\mid X=x]$$

(b)

b.png

Figure 1: empirical and true errors, averaged across the T, as a function of n

ניתן לראות כי ככל ש n גדל השגיאה האמפירית גדלה, וכן השגיאה האמיתית קטנה

כאשר מספר הדגימות מצומצם, סביר ש ERM יחזיר היפותזה בעלת שגיאה קרובה מאוד ל 0, ולכן יש שונות גבוהה באינטרוולים ש ERM יחסיר וניתן לראות זאת עפ"י שגיאת האמת (בכחול). ככל ש n גדל אנו נצפה ש ERM יחזיר היפותזה קרובה מאוד להיפותזה בעלת שגיאת האמת הקטנה ביותר, שכן הערכה להתפלגות האמיתית גדלה (ברמת וודאות). ניתן לראות כי אכן זה המצב שכן שגיאת האמת עבור n גדולים (קטנה). סביר להעריך כי עבור n גדולים יותר ויותר המרחק בין השגיאה האמפירית לאמיתית יצטמצם.

(c)

`c.png`

Figure 2: empirical and true errors as a function of k

ניתן לראות כי ככל ש k השגיאה האמפירית קטנה (אם כי ניתן לראות כי עד $k = 3$ ישנה ירידה חדה בשגיאה ואחריה ירידה מתונה) כלומר $k^* = 10$

ניתן לייחס זאת ל overfitting

אנו שמים לב כי לקיחת היפותזה שלk^* לא בהכרח טובה שכן עבור $k^* > k = 3$ מתקבלת שגיאת אמת קטנה ביותר - כמצופה (השגיאה הטובה ביותר נמצאת במחלקה $\mathcal{H}_3$).

(d)

`d.png`

Figure 3: holdout

The best empirical k : 3

The best hyposesis that found represent as a set of intervals:

$$[(0.000279, 0.199099), (0.399137, 0.599962), (0.800406, 0.995437)]$$

אנו רואים כי ההיפותזה עבור $k = 3$ מקבלת את השגיאה האמפירית הקטנה ביותר עבור דגימת ה test ולכן לפי שיטת holdout אנו נבחר בהיפותזה לעיל.

נשים לב כי ההיפותזה שחזרה קרובה מאוד להיפותזה האופטימלית :

$$[(0, 0.2), (0.4, 0.6), (0.8, 1)]$$