הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ג – 14 במרץ 2023

מרצה: ד"ר יונתן בלינקוב

מבחן מסכם מועד ב' – פתרון חלקי

שימו לב: הפתרונות המופיעים כאן הם חלקיים בלבד ומובאים בשביל לעזור לכם בתהליך הלמידה.

ייתכנו כאן חוסרים / ליקויים / טעויות של ממש.

הנחיות הבחינה:

- **משך הבחינה:** שלוש שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- מחשבון: מותר.
- כלי כתיבה: עט בלבד.
- יש לכתוב את התשובות **על גבי שאלון זה**.
- מותר לענות בעברית או באנגלית.
- הוכחות והפרכות צריכות להיות פורמליות.
- קריאוּת:
  - תשובה בכתב יד לא קריא – **לא תיבדק**.
  - בשאלות רב-ברירה – הקיפו את התשובות בבירור. סימונים לא ברורים יביאו לפסילת התשובה.
  - לא יתקבלו ערעורים בנושא.
- במבחן 14 עמודים ממוספרים סה"כ, כולל עמוד שער זה שמספרו 1 ושלושה עמודי טיוטה בסוף הגיליון.
- נא לכתוב רק את המבוקש ולצרף הסברים קצרים עפ"י ההנחיות.
- **בתום המבחן יש להגיש את שאלון זה בלבד.**

בהצלחה!

חלק א' – שאלות פתוחות [82 נק']

שאלה 1: השפעה של דוגמה יחידה על פעולת מסווגים [24 נק']

נתון סֵט אימון עם $m \geq 10$ דוגמאות דו-ממדיות ותיוגים בינאריים, משמע לכל $i = 1, \ldots, m$ מתקיים $\mathbf{x}_i \in \mathbb{R}^2, y_i \in \{-1,1\}$.

לומדים שני מסווגים:

- בשלב הראשון: לומדים מסווג על סט האימון המקורי ומחשבים את הסיווגים על כל הדוגמאות.
- בשלב השני:
  - מסירים דוגמת אימון אחת שרירותית כלשהי.
  - מאמנים מסווג חדש על סט האימון המעודכן, ומחשבים בעזרתו את הסיווג על כל הדוגמאות הנותרות.

עבור כל אלגוריתם למידה, סמנו האם הסיווגים שהמסווג החדש יחזה על דוגמאות האימון הנותרות זהים בהכרח לאלה של המסווג המקורי על דוגמאות אלה.

הסבירו בקצרה את תשובותיכם (2-4 משפטים בכל סעיף).

הניחו שאין צעדים אקראיים או שגיאות נומריות בריצת האלגוריתמים (בעיות קמורות מתכנסות לפתרון האנליטי במדויק).

א. Soft-SVM ליניארי לא הומוגני עם $\lambda = 10^{-1}$ בהנחה שהדאטה המקורי פריד ליניארית.

הסיווגים על דוגמאות האימון הנותרות זהים בהכרח? כן / **לא**

הסבר: ייתכן שהנקודה שהוסרה הייתה ב-margin violation גבוה ו"דחפה" את המפריד לכיוונה. לאחר הסרתה, המפריד יכול לזוז בצורה **משמעותית** ונקודה שסווגה לפני כן באופן נכון יכולה כעת להיות מסווגת באופן שגוי.

ב. Soft-SVM ליניארי לא הומוגני עם $\lambda \to 0$ בהנחה שהדאטה המקורי פריד ליניארית.

הסיווגים על דוגמאות האימון הנותרות זהים בהכרח? **כן** / לא

הסבר: בגבול $\lambda \to 0$ מקבלים התנהגות של Hard SVM. דאטה שהיה פריד ליניארית נשאר כך לאחר הסרת דוגמה.

ג. ID3 המשתמש באנטרופיה ועוצר בעומק מירבי 4. הסיווגים על דוג' האימון הנותרות זהים בהכרח? כן / **לא**

הסבר: חישובי האנטרופיה וה-IG כבר בשלב הראשון משתנים ולכן ייתכן שייבחר כלל פיצול אחר בשורש וייבנה עץ אחר לחלוטין.

ד. kNN עם $k = 3$ (דוגמה לא נחשבת שכנה של עצמה), כאשר ידוע שלשלושת השכנים הקרובים ביותר לדוגמה שהוסרה יש תיוג זהה לתיוג שלה. הסיווגים על דוגמאות האימון הנותרות זהים בהכרח? כן / **לא**

הסבר: שכנות היא לא יחס טרנזיטיבי. דוגמה נגדית (מסירים ב-6, הסיווג על זו ב-4 משתנה):

שאלה 2: Generative models [26 נק']

תזכורת: פונק' הצפיפות של התפלגות $U[a,b]$, אחידה ורציפה על הקטע הסגור $[a,b]$, היא

$$f(z)=\frac{1}{b-a}\mathbb{I}[a\le z\le b]=\begin{cases}\frac{1}{b-a}, & a\le z\le b\\ 0, & \text{otherwise}\end{cases}.$$

א. [5 נק'] **מתרגיל בית:** נתון משתנה אקראי $X\sim U[0,\theta]$ עבור $\theta>0$ לא ידוע.

נתון מדגם אקראי S של m דגימות, $S=\{x_1,\dots,x_m\}\subset\mathbb{R}_{\ge 0}$, שנדגמו מהמשתנה האקראי באופן i.i.d.

הוכיחו שמשערך ה-MLE שמוגדר בתור $\hat{\theta}_{\text{MLE}}\triangleq\operatorname{argmax}_\theta\underbrace{\Pr[S;\theta]}_{\text{likelihood}}$ הוא $\hat{\theta}_{\text{MLE}}=\max_{i\in[m]}x_i$.

תשובה:

פונקציית ה-likelihood היא $\Pr[S;\theta]=\prod\Pr[x_i;\theta]=\prod\frac{1}{\theta-0}\mathbb{I}[0\le z\le\theta]=\frac{1}{\theta^m}\mathbb{I}[\forall i: 0\le x_i\le\theta]$.

אם אפילו אחת הדוגמאות $x_i>\theta$ הפונק' היא אפס. בתחום $\theta\ge\max_i x_i$ הפונק' לא אפס והיא יורדת.

כדי למקסם אותה נבחר $\hat{\theta}_{\text{MLE}}=\max_i x_i$.

ב. [6 נק'] בנוסף על הנתונים שבסעיף הקודם, בסעיף זה בלבד נתון שמתקיים $\theta\sim U[10,20]$.

מצאו (והוכיחו) את משערך ה-MAP לפי כלל הנתונים: $\hat{\theta}_{\text{MAP}}\triangleq\operatorname{argmax}_\theta\Pr[\theta|S]=\operatorname{argmax}_\theta\underbrace{\Pr[S|\theta]}_{\text{likelihood}}\underbrace{\Pr[\theta]}_{\text{prior}}$.

תשובה:

מתקבל $\Pr[S;\theta]\Pr[\theta]=\frac{1}{\theta^m}\mathbb{I}\left[\theta\ge\max_i x_i\right]\frac{1}{10}\mathbb{I}[10\le\theta\le 20]$.

ושוב הפונק' יורדת בטווחים בהם היא אינה אפס. ולכן נבחר $\hat{\theta}_{\text{MAP}}=\max\left(10,\max_i x_i\right)$

בסעיף הבא מרחב הדוגמאות הוא $\mathcal{X}=\mathbb{R}^2_{\geq 0}$ (הרביע החיובי) ומרחב התיוגים הוא $\mathcal{Y}=\{-1,+1\}$.

בהינתן מדגם אימון $\{(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_m,y_m)\}\subset(\mathcal{X}\times\mathcal{Y})$, נרצה ללמוד מסווג בינארי.

<u>תהליך הלמידה:</u>

i. נניח את הנחת Naïve Bayes (NB).

ii. נמדל את בעיית הסיווג בעזרת Uniform NB, משמע, $(X[j]|Y=k)\sim U\left[0,\theta_k[j]\right]$, כאשר $k\in\{-1,+1\}, j\in\{1,2\}$.

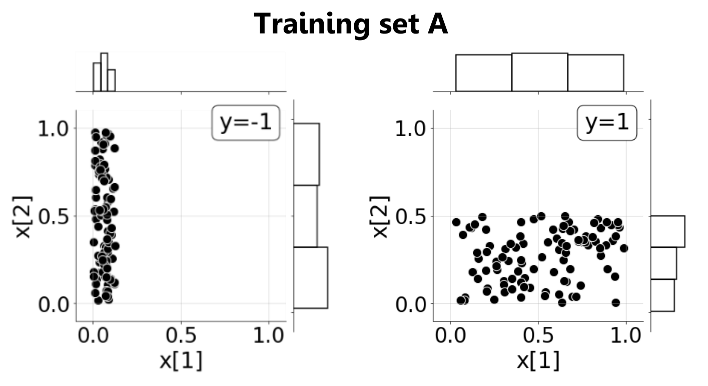
iii. נשערך את ארבעת הפרמטרים בעזרת MLE, משמע $\hat{\theta}_{+1}=\begin{bmatrix}\max_{i:y_i=+1} x_i[1]\\ \max_{i:y_i=+1} x_i[2]\end{bmatrix}$ ו-$\hat{\theta}_{-1}=\begin{bmatrix}\max_{i:y_i=-1} x_i[1]\\ \max_{i:y_i=-1} x_i[2]\end{bmatrix}$.

iv. בהמשך לכל ההנחות לעיל, נבנה כלל החלטה הסתברותי $\hat{y}(\mathbf{x})=\mathrm{argmax}_{y\in\{-1,+1\}}\Pr(\mathbf{x};y)$.

כעת נתונים שלושה מדגמי אימון, כל אחד מהתפלגות שונה ומכיל 100 דוגמאות חיוביות ו-100 שליליות. המדגמים מופיעים בתרשימים הבאים (הדוגמאות מִכּל תיוג מופיעות בנפרד ביחד עם ההיסטוגרמות השוליות המתאימות).

לכל מדגם (בנפרד), מבצעים את תהליך הלמידה המתואר לעיל.

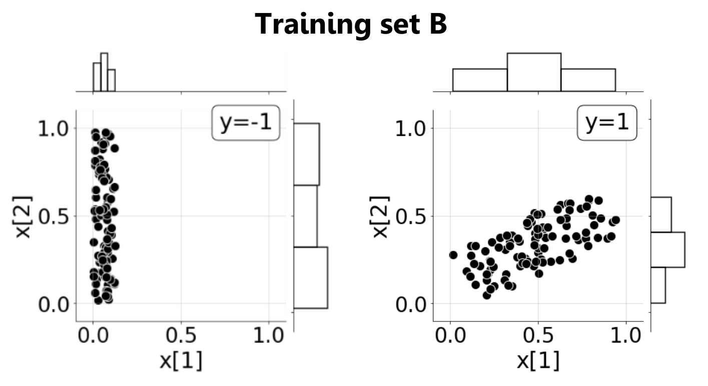
ג. [15 נק'] לכל מדגם, ענו על השאלות ביחס לתהליך הלמידה שלו. התשובות אמורות להיות ברורות מהגרפים.



האם נראה שהנחת NB מתקיימת? כן / **לא**

איך תסווג הנקי' $(0,0)$? $+1$ / -1 / **לא ניתן לדעת**

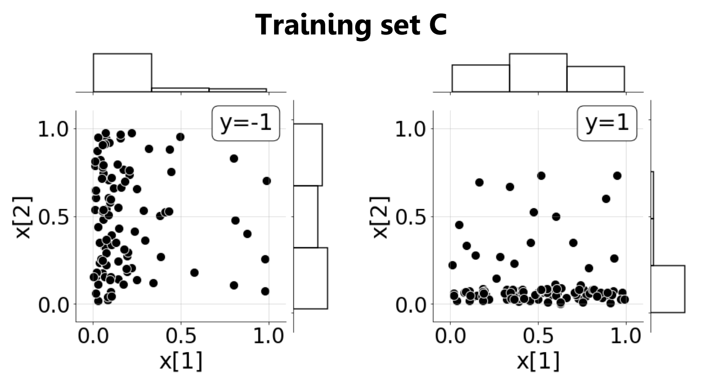
האם דיוק האימון יהיה מעל 70%? כן / **לא**



האם נראה שהנחת NB מתקיימת? כן / **לא**

איך תסווג הנקי' $(0,0)$? $+1$ / -1 / **לא ניתן לדעת**

האם דיוק האימון יהיה מעל 70%? כן / **לא**



האם נראה שהנחת NB מתקיימת? כן / **לא**

איך תסווג הנקי' $(0,0)$? $+1$ / -1 / **לא ניתן לדעת**

האם דיוק האימון יהיה מעל 70%? כן / **לא**

שאלה 3: Multi-Layer Perceptron (MLP) and VC dimension [32 נק']

קראו היטב את הנתונים הבאים.

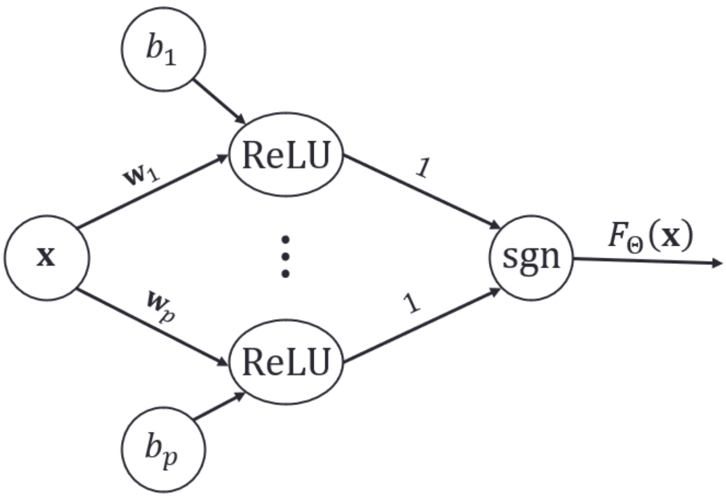
בשאלה זו מרחב הנתונים הוא $\mathcal{X} = \mathbb{R}^2, \mathcal{Y} = \{-1, +1\}$.

נגדיר מחלקה $\mathcal{H}$ של רשתות MLP עם שכבה חבויה אחת ברוחב $p \in \mathbb{N}$ (היפר-פרמטר), אקטיבציות ReLU ופלט <u>בינארי</u> יחיד.

בכל הרשתות במחלקה, המשקלים של השכבה השנייה קבועים להיות 1 וללא bias.

נאמר שאוסף פרמטרים Θ הוא **<u>חוקי</u>**, אם המשקלים בו <u>אי-שליליים</u> (אילוץ זה לא כולל את רכיבי ה-bias).

נתאר את הפונקציה המתקבלת $F_\Theta: \mathbb{R}^2 \to \{-1, +1\}$ בצורה גרפית ובצורה פורמלית:



$$F_\Theta(\mathbf{x}) = \operatorname{sgn}\left(\sum_{t=1}^{p} \operatorname{ReLU}(\mathbf{w}_t^\top \mathbf{x} + b_t)\right),$$

where:

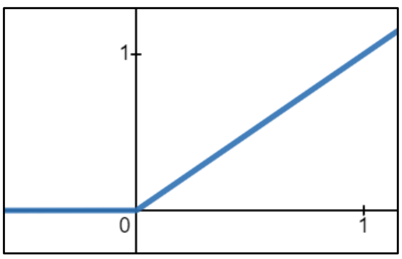
$$\Theta = (\mathbf{w}_1, \dots, \mathbf{w}_p, b_1, \dots, b_p),$$
$$\mathbf{w}_1, \dots, \mathbf{w}_p \in \mathbb{R}^2_{\geq 0},$$
$$b_1, \dots, b_p \in \mathbb{R}.$$

וכמו כן,

נגדיר: $\operatorname{sgn}(z) = \begin{cases} -1, & z \leq 0 \\ +1, & z > 0 \end{cases}$

(כך שלא מתקבל אפס לשום קלט)

תזכורת: $\operatorname{ReLU}(z) = \begin{cases} 0, & z \leq 0 \\ z, & z > 0 \end{cases}$



סימון: יהיו שני קלטים $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^2$. נסמן $\mathbf{x}_j \succcurlyeq \mathbf{x}_i$ אם ורק אם $\mathbf{x}_i[1] \geq \mathbf{x}_j[1] \wedge \mathbf{x}_i[2] \geq \mathbf{x}_j[2]$.

א. [8 נק'] הוכיחו: ברשתות שהגדרנו, לכל רוחב p ולכל Θ חוקי, אם $\mathbf{x}_j \succcurlyeq \mathbf{x}_i$ אזי $F_\Theta(\mathbf{x}_i) \geq F_\Theta(\mathbf{x}_j)$.

הוכחה (לרשותכם טיוטה בסוף הגיליון):

נניח $\mathbf{x}_i \succcurlyeq \mathbf{x}_j$.

מתקיים $\underbrace{\overbrace{\mathbf{w}_t[1]}^{\geq 0}\mathbf{x}_j[1] + \overbrace{\mathbf{w}_t[2]}^{\geq 0}\mathbf{x}_j[2] + b_t}_{=\mathbf{w}_t^\top \mathbf{x}_i + b_t} \geq \underbrace{\overbrace{\mathbf{w}_t[1]}^{\geq 0}\mathbf{x}_i[1] + \overbrace{\mathbf{w}_t[2]}^{\geq 0}\mathbf{x}_i[2] + b_t}_{=\mathbf{w}_t^\top \mathbf{x}_i + b_t}$:$\forall t \in [p]$.

ולכן גם $\forall t \in [p]$ מתקיים $\mathrm{ReLU}(\mathbf{w}_t^\top \mathbf{x}_j + b_t) \geq \mathrm{ReLU}(\mathbf{w}_t^\top \mathbf{x}_i + b_t)$ ובסה"כ $F_\Theta(\mathbf{x}_i) \geq F_\Theta(\mathbf{x}_j)$.

ב. [8 נק'] נגדיר את ה-XOR dataset: $S = \left\{\left(\underbrace{(0,1)}_{\mathbf{x}_1}, \underbrace{+1}_{y_1}\right), \left(\underbrace{(1,0)}_{\mathbf{x}_2}, \underbrace{+1}_{y_2}\right), \left(\underbrace{(0,0)}_{\mathbf{x}_3}, \underbrace{-1}_{y_3}\right), \left(\underbrace{(1,1)}_{\mathbf{x}_4}, \underbrace{-1}_{y_4}\right)\right\}$.

הוכיחו שלכל רוחב p ולכל Θ חוקי, לא ניתן להגיע לשגיאת אימון אפס על S ע"י $F_\Theta \in \mathcal{H}$ (כדאי להיעזר בסעיף הקודם).

הבהרה: כל הסעיפים עוסקים ב-capacity של המחלקה ולא במציאת דרכי אימון יעילות.

הוכחה:

מצד אחד מתקיים $y_2 > y_4$, $\mathbf{x}_4 \succcurlyeq \mathbf{x}_2$. מצד שני $\hat{y}_4 = F_\Theta(\mathbf{x}_4) \geq F_\Theta(\mathbf{x}_2) = \hat{y}_2$ לכל Θ חוקי.

ג. [8 נק'] בסעיף זה נבנה רשת (מתוך $\mathcal{H}$) ברוחב $p = 1$.

יהיו $n \geq 2$ קלטים $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^2_{\geq 0}$ שונים ומנורמלים ברביע החיובי (משמע $\forall i$ מתקיים $\|\mathbf{x}_i\|_2 = 1$ וגם $\forall j \neq i: \mathbf{x}_i \neq \mathbf{x}_j$).

הראו שקיימת השמה <u>חוקית</u> $\Theta = (\underbrace{\mathbf{w}_1}_{\in\mathbb{R}^2_{\geq 0}}, \underbrace{b_1}_{\in\mathbb{R}})$ שמקיימת $F_\Theta(\mathbf{x}_1) = +1$ וגם $F_\Theta(\mathbf{x}_2) = F_\Theta(\mathbf{x}_3) = \cdots = F_\Theta(\mathbf{x}_n) = -1$.

משמע, הציעו השמה כזאת (שתלויה ב-$(\mathbf{x}_1, \dots, \mathbf{x}_n)$) והוכיחו שהיא מקיימת את הנדרש.

<u>רמז</u>: ניתן להיעזר בזהות האלגברית $\mathbf{u}^\top \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\|\cos\angle(\mathbf{u}, \mathbf{v})$.

הוכחה (לרשותכם טיוטה בסוף הגיליון):

נבחר $\mathbf{w}_1 = \mathbf{x}_1$.

$$\forall i: \mathrm{ReLU}(\mathbf{w}_1^\top \mathbf{x}_i + b_1) = \mathrm{ReLU}(\mathbf{x}_1^\top \mathbf{x}_i + b_1) = \begin{cases} 1 + b_1, & i = 1 \\ \cos\angle(\mathbf{x}_1, \mathbf{x}_i) + b_1, & i = 2, \dots, \mathrm{n} \end{cases}$$

צריך לוודא שמתקיים $1 + b_1 > 0$ וגם $\cos\angle(\mathbf{x}_1, \mathbf{x}_i) + b_1 \leq 0$. משמע $\underbrace{\cos\angle(\mathbf{x}_1, \mathbf{x}_i)}_{\in[0,1), \forall i \geq 2} \leq -b_1 < 1$.

לכן נבחר $b_1 = -\max_{\mathrm{i}=2,\dots,\mathrm{n}} \cos\angle(\mathbf{x}_1, \mathbf{x}_i)$.

בהמשך ניתן להיעזר בסעיף הקודם גם מבלי לפתור אותו (נדרשות התאמות כי בסעיף הקודם $p = 1$ בלבד).

ד. [8 נק'] נסמן ב-$\mathcal{H}_p$ את מחלקת הרשתות מ-$\mathcal{H}$ שהן ברוחב $p \in \mathbb{N}$ (משמע, מתקיים $\mathcal{H}_p \subset \mathcal{H}$).

מבין האפשרויות הבאות, בחרו את <u>החסם התחתון ההדוק ביותר</u> שתוכלו להוכיח עבור $p \geq 2$:

(i) $\mathrm{VCdim}(\mathcal{H}_p) \geq 2$ (ii) $\mathrm{VCdim}(\mathcal{H}_p) \geq \ln p$ (iii) $\boldsymbol{VCdim(\mathcal{H}_p) \geq p}$

הוכיחו את החסם התחתון שבחרתם.

הוכחה: נבחר p נקודות שונות מנורמלות ברביע הראשון (הבחירה עצמה לא משנה כאן).

אינטואיציה: אפשר להסיק מהסעיף הקודם, שניתן ליצור נוירון בודד חבוי שנדלק רק עבור דוגמה אחת.

תהא השמה כלשהי $y_1, \dots, y_p$.

בדומה לסעיף הקודם, לכל $t \in [p]$:

אם $y_t = 1$: נבחר $\mathbf{w}_t = \mathbf{x}_t$ וגם $b_t = -\max_{i \neq t} \cos\angle(\mathbf{x}_t, \mathbf{x}_i)$. מתקיים $\mathrm{ReLU}(\mathbf{w}_t^\top \mathbf{x}_i + b_t) > 0 \Leftrightarrow i = t$.

יתקיים בהכרח $F_\Theta(\mathbf{x}_t) = 1$ (כל שאר ה-ReLU הם אפסים בהתאם לבנייה).

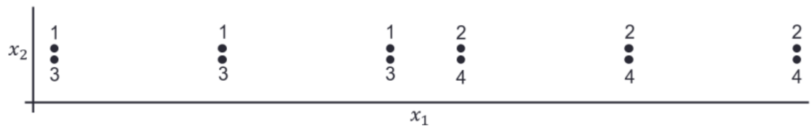
אם $y_t = -1$: נבחר $\mathbf{w}_t = \mathbf{0}_d, b_t = 0$. מתקיים לכל i: $\mathrm{ReLU}(\mathbf{w}_t^\top \mathbf{x}_i + b_t) = 0$.

יתקיים בהכרח $F_\Theta(\mathbf{x}_t) = 0$ (כל שאר ה-ReLU הם אפסים בהתאם לבנייה).

חלק ב' – שאלות רב-ברירה [18 נק']

הערות בדיקה: בשתי השאלות הראשונות ירדו 2 נק' (מתוך 6) על כל סימון שגוי. בשלישית ניתנה נק' 1 לסימון של d או e.

א. לפניכם סט אימון דו-ממדי עם 4 מחלקות ו-3 דוגמאות מכל מחלקה (התיוג כתוב מעל/מתחת הדוגמאות).



מבין מודלי ה-multiclass הבאים, סמנו את כֹּל אלה שצפויים להגיע לדיוק אימון של 100% על הדאטה לעיל.

a. 1-nearest-neighbor (חוזה את התיוג של השכן הקרוב ביותר לפי מרחק אוקלידי, דוג' לא נחשבת שכנה של עצמה).

b. עץ החלטה בעומק מירבי 3 (הפרדיקציה של כל עלה נקבעת לפי רוב דוגמאות האימון שבתוכו).

c. מודל one-vs-one עם decision stump (עץ בעומק 1) כמודל בסיס.

d. מודל one-vs-all עם decision stump (עץ בעומק 1) כמודל בסיס.

ב. נגדיר אלגוריתם Random Forest פשוט:

Random Forest(S, k, max_depth, min_samples_split**):**
For i=1 to k:
S' = Sample $\sqrt{d}$ features out of the original d features in S (keeping all samples)
h_i = ID3(S', max_depth, min_samples_split, criterion="entropy")
Return $H(x) = \frac{1}{k}\sum_{i=1}^{k} h_i(x)$

אילו מבין הבחירות האלגוריתמיות הבאות צפויות להפחית את ה-Variance של המסווג הכולל שנלמד (H)?

סמנו את כֹּל התשובות הנכונות (השאלה אינה עוסקת במקרי קצה אלא במקרה הסביר).

a. הגדלת k (מספר העצים ביער).

b. הגדלת max_depth (העומק המירבי המותר).

c. הגדלת min_samples_split (מספר הדוגמאות המינימלי הנדרש לפיצול של צומת).

d. נירמול מקדים של הדאטה בשיטת min-max.

e. נירמול מקדים של הדאטה בשיטת standardization (Z-score).

ג. נתון מדגם $S=\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ של דוגמאות ב-$\mathbb{R}^d$ ותיוגים בינאריים (±1). נתונה פונקציית מיפוי כלשהי $\boldsymbol{\phi}: \mathbb{R}^d \to \mathbb{R}^{10}$.

נתון שקיימים $\widehat{\mathbf{w}} \in \mathbb{R}^{10}$ וקבוע $c > 0$ המקיימים $\forall i \in [m]: y_i \widehat{\mathbf{w}}^\top \boldsymbol{\phi}(\mathbf{x}_i) \geq c$.

נגדיר בעיית אופטימיזציה בעזרת hinge loss על S: $\underset{\mathbf{w}}{\operatorname{argmin}} \left(\frac{1}{m}\sum_i \max\{0,\ 1 - y_i \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i)\}\right)$.

אם נמצא מינימום לוקאלי $\overline{\mathbf{w}}$ של הבעיה, האם שגיאת ה-0-1 של $\overline{\mathbf{w}}$ (על S, לאחר המיפוי ϕ) היא בהכרח אפס?

סמנו את התשובה הנכונה.

a. כן.

b. רק אם הפונקציה ϕ ליניארית.

c. רק אם הפונקציה ϕ לא ליניארית.

d. רק אם $c = 1$.

e. רק אם $c \geq 1$.

f. לא, כי חסר גורם רגולריזציה.