



מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ג – 14 במרץ 2023

מרצה: ד"ר יונתן בלינקוב

מבחן מסכם מועד ב'

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.
- קריאות:
- 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 הגליון.
- 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 סט אימון עם $m \geq 10$ דוגמאות דו-ממדיות ותיגים בינאריים, משמע לכל $i = 1, \dots, m$ מתקיים $y_i \in \{-1, 1\}$, $x_i \in \mathbb{R}^2$.

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 הנתרות.

עבור כל אלגוריתם למידה, סמנו האם הסיווגים שהמסווג החדש יחזה על דוגמאות האימון הנתרות 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).

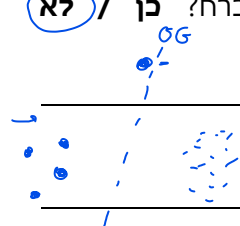
א. Soft-SVM ליניארי לא הומוגני עם $\lambda = 10^{-1}$ בהנחה שהדאטה המקורי פריד ליניארית.

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

הסבר: הסרת הנק' עלולה לשנות את המסלול של הmargin → small Margin → large Norm

אם כן, הן יסווג אותה נק' אחרת כלשהי.

כאשר כוונתנו למשחקי margin, Norm,



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

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

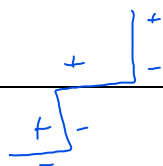
הסבר: הדאטה פריד ליניארית ולכן יבנו מסלול שיסווג את כל המבנים.

אם כוונתנו למשחקי margin, Norm,

hard SVM

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

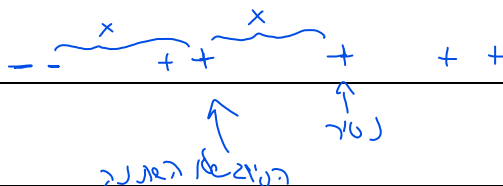
הסבר: $m \geq 10$



הסבר: יהיה להוביל לפעולת החישוב האנליטית, ובצדק
אלה הם הכללים והכללים הם שונים, אך הם

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:

שאלה 2: Generative models [26 נק']

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

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

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, $S = \{x_1, \dots, x_m\} \subset \mathbb{R}_{\geq 0}$, שנדגמו מהמשתנה האקראי באופן i.i.d.

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

תשובה:

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:

בסעיף הבא מרחב הדוגמאות הוא $\mathcal{X} = \mathbb{R}_{\geq 0}^2$ (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.

תהליך הלמידה:

i. נניח את הנחת Naïve Bayes (NB).

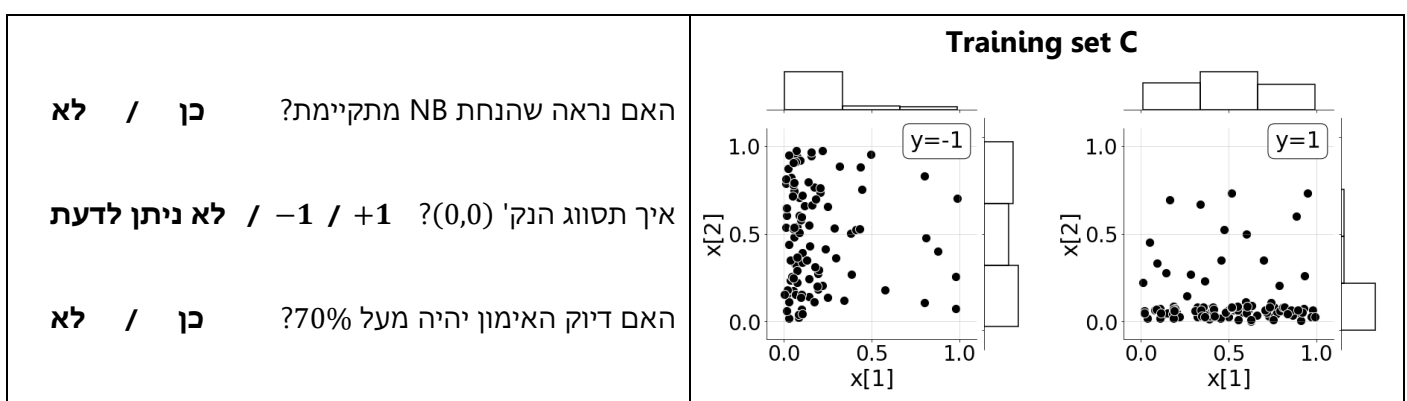
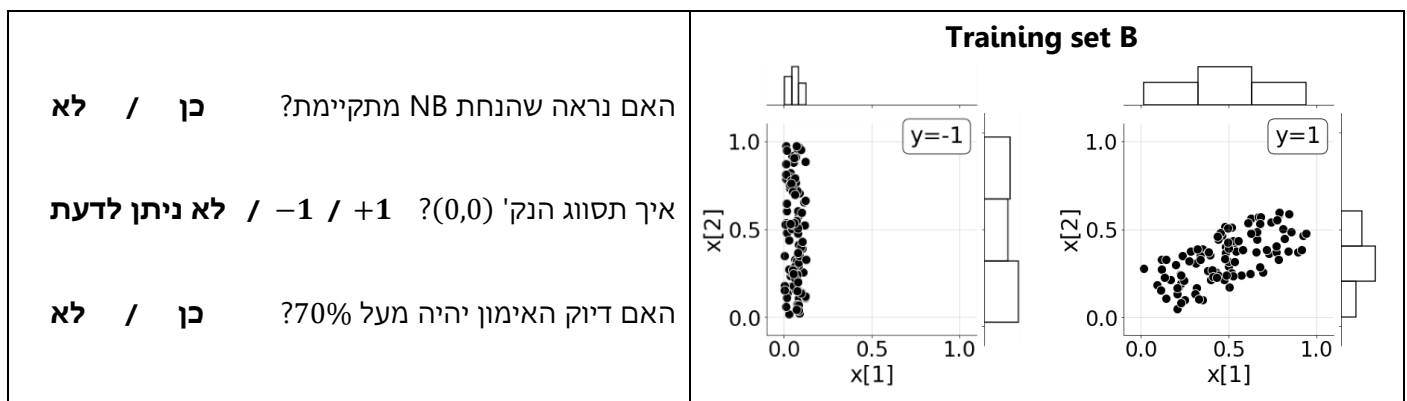
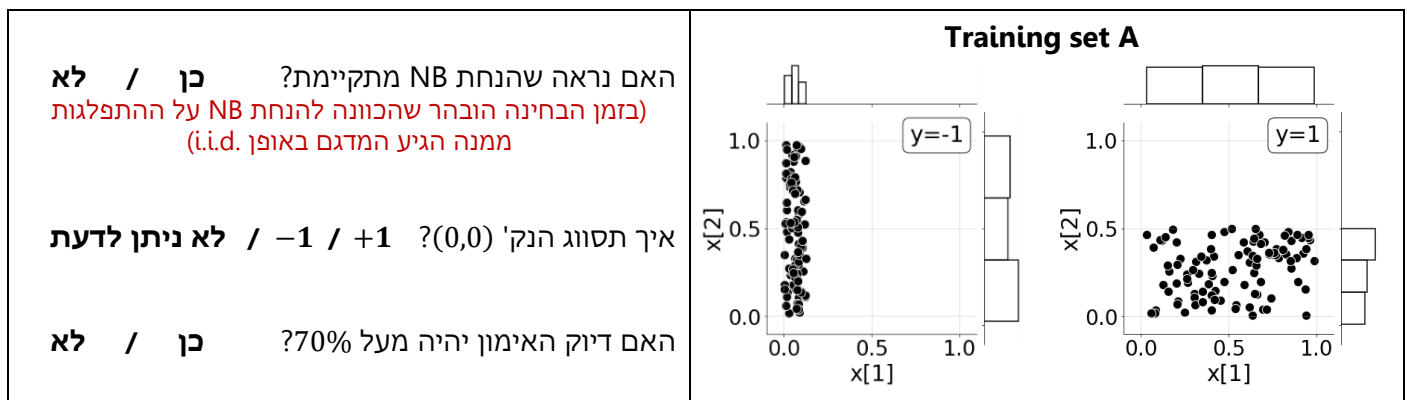
ii. נמדל את בעיות הסיווג בעזרת Uniform NB, משמע, $X[j]|Y=k) \sim U[0, \theta_k[j]]$, כאשר $j \in \{1, 2\}, k \in \{-1, +1\}$.

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

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

כעת נתונים שלושה מדגמי אימון, כל אחד מהתפלגות שונה ומכיל 100 דוגמאות חיוביות ו-100 שליליות. המדגמים מופיעים בתרשימים הבאים (הדוגמאות מכל 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.



שאלה 3: Multi-Layer Perceptron (MLP) and VC dimension [23 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), אקטיביציות ReLU ופלט בינארי יחיד.

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

נאמר שאוסף פרמטרים θ הוא **חוקי**, אם המשקלים בו אי-שליליים (אילוץ זה לא כולל את רכיבי ה-bias).נתאר את הפונקציה המתקבלת $F_\theta: \mathbb{R}^2 \rightarrow \{-1, +1\}$ בצורה גרפית ובצורה פורמלית:

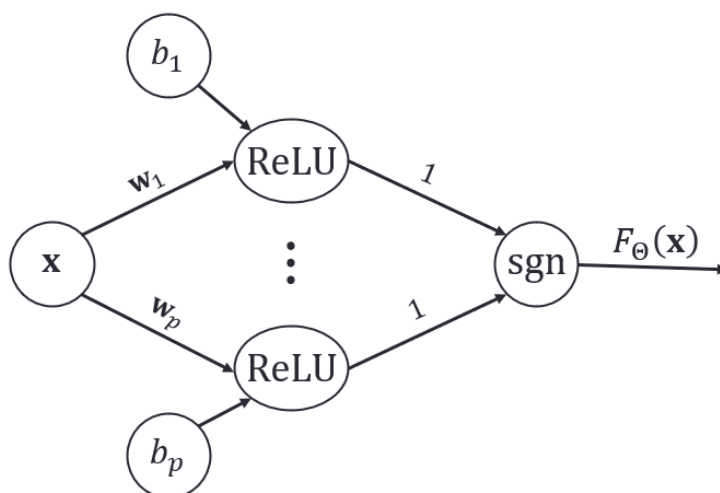
$$F_\theta(\mathbf{x}) = \text{sgn} \left(\sum_{t=1}^p \text{ReLU}(\mathbf{w}_t^\top \mathbf{x} + b_t) \right),$$

where:

$$\theta = (\mathbf{w}_1, \dots, \mathbf{w}_p, b_1, \dots, b_p),$$

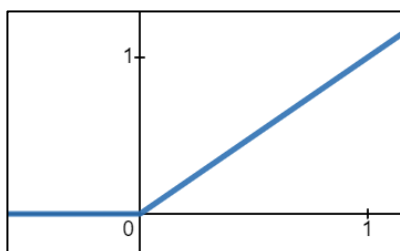
$$\mathbf{w}_1, \dots, \mathbf{w}_p \in \mathbb{R}_{\geq 0}^2,$$

$$b_1, \dots, b_p \in \mathbb{R}.$$



וכמו כן,

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



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

(כך שלא מתקבל אפס לשום קלט)

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

א. [8 נק'] הוכיחו: ברשתות שהגדרנו, לכל רוחב p ולכל θ חוקי, אם $x_i \succcurlyeq x_j$ אזי $F_\theta(x_i) \geq F_\theta(x_j)$.

הוכחה (לרשותכם טיוטה בסוף הגיליון):

$$F_\theta(x_i) = \text{sgn}\left(\sum_{k=1}^p \text{Relu}(w_k^T x_i + b_k)\right) \stackrel{(*)}{=} \text{Sign}\left(\sum_{k=1}^p \text{Relu}(w_k^T x_i + b_k)\right)$$

$$x_i \succcurlyeq x_j \Rightarrow \begin{matrix} \uparrow \\ w \in \mathbb{R}_{\geq 0}^2 \end{matrix} w^T x_i \geq w^T x_j \Rightarrow \begin{matrix} \uparrow \\ b \in \mathbb{R} \end{matrix} (w^T x_i + b) \geq (w^T x_j + b) \Rightarrow$$

$$\Rightarrow \text{Relu}(x_i) \geq \text{Relu}(x_j) \Rightarrow$$

$$\Rightarrow \sum \text{Relu}(x_i) \geq \sum \text{Relu}(x_j) \Rightarrow$$

$$\Rightarrow \text{Sign}\left(\sum \dots x_i \dots\right) \geq \text{Sign}\left(\sum \dots x_j \dots\right) \Rightarrow$$

$$\Rightarrow F_\theta(x_i) \geq F_\theta(x_j)$$

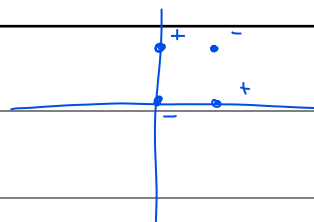
$$S = \left\{ \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} +1 \end{pmatrix} \right), \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} +1 \end{pmatrix} \right), \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \end{pmatrix} \right), \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \end{pmatrix} \right) \right\}$$

ב. [8 נק'] נגדיר את ה-XOR dataset:

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).

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

הוכחה:



$$\forall F_\theta(x_i) = y_i$$

$$\text{כדי כי נוקט } \Leftarrow \text{היה } F_\theta \in \mathcal{H} \text{ נכח}$$

$$F_\theta(x_1) = F_\theta(x_2) = 1$$

סליל

$$F_\theta(x_3) = F_\theta(x_4) = -1$$

$$1 = x_4[0] \geq x_1[0] = 0$$

$$x_4 \geq x_1 \text{ כי}$$

יור כי יור

$$\text{כדי כי } -1 = F_\theta(x_4) \geq F_\theta(x_1) = 1 \text{ סתירה} \Rightarrow \text{לא ניתן להגיע לאימון אפס}$$

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

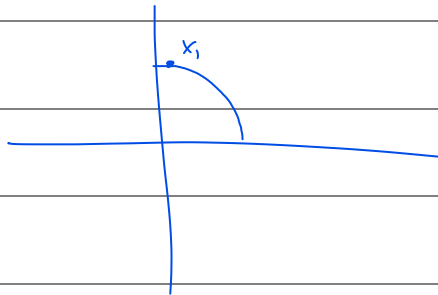
יהיו $n \geq 2$ קלטים $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}_{\geq 0}^2$ שונים ומנורמלים ברביע החיובי (משמע $\forall i$ מתקיים $\|\mathbf{x}_i\|_2 = 1$ וגם $\forall j \neq i: \mathbf{x}_i \neq \mathbf{x}_j$).

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

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

במז: ניתן להיעזר בזהות האלגברית $\mathbf{u}^T \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos \angle(\mathbf{u}, \mathbf{v})$.

הוכחה (לרשותכם טיוטה בסוף הגיליון):



$$\sqrt{(\cdot)^2}$$

$$\omega_i = \mathbf{x}_i \Rightarrow \forall i: \text{ReLU}(\omega_i^T \mathbf{x}_i + b_i) = \text{ReLU}(\mathbf{x}_i^T \mathbf{x}_i + b_i) = \begin{cases} b_i & ; i=1 \\ \cos \angle(\mathbf{x}_i, \mathbf{x}_i) & ; i=2, \dots, n \end{cases}$$

$$\cos \angle(\mathbf{x}_i, \mathbf{x}_i) + b_i \leq 0 \quad \text{אם } b_i > 0 \quad \text{אם } b_i \leq -1$$

$$b_i \leq -\cos \angle(\mathbf{x}_i, \mathbf{x}_i)$$

$$b_i \geq -1$$

$$b$$

$$\cos \angle(\mathbf{x}_i, \mathbf{x}_i) \leq -b < 1$$

$$b_i = -\max_{i=2, \dots, n} \cos \angle(\mathbf{x}_i, \mathbf{x}_i)$$

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 החסם התחתון ההדוק ביותר שתוכלו להוכיח עבור $p \geq 2$:

(i) $\text{VCdim}(\mathcal{H}_p) \geq 2$

(ii) $\text{VCdim}(\mathcal{H}_p) \geq \ln p$

(iii) $\text{VCdim}(\mathcal{H}_p) \geq p$

הוכיחו את החסם התחתון שבחרתם.

הוכחה:

תהי p ו- y_1, \dots, y_p חסמי

נסתכל ב- \mathcal{H}_p ונראה ש- $\text{VCdim}(\mathcal{H}_p) \geq p$

נבנה רשתות f_t עבור $t \in [p]$ כך ש-

$f_t(x_i) = y_i$ עבור $i \leq t$ ו- $f_t(x_i) = 0$ עבור $i > t$

נראה ש- $f_t \in \mathcal{H}_p$ עבור כל t

נבנה רשתות f_t עבור $t \in [p]$ כך ש-

$f_t(x_i) = y_i$ עבור $i \leq t$ ו- $f_t(x_i) = 0$ עבור $i > t$

נראה ש- $f_t \in \mathcal{H}_p$ עבור כל t

נבנה רשתות f_t עבור $t \in [p]$ כך ש-

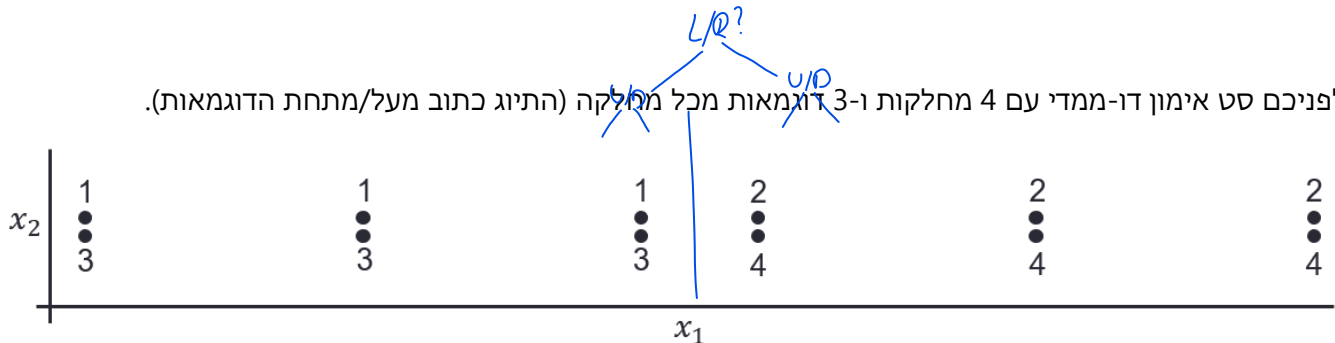
$f_t(x_i) = y_i$ עבור $i \leq t$ ו- $f_t(x_i) = 0$ עבור $i > t$

נראה ש- $f_t \in \mathcal{H}_p$ עבור כל t

חלק ב' – שאלות רב-ברירה [18 נק']

בשאלות הבאות סמנו את התשובות המתאימות (לפי ההוראות). בחלק זה אין צורך לכתוב הסברים.

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 כל אלה שצפויים להגיע לדיוק אימון של 100% על הדאטה לעיל.

a. 1-nearest-neighbor (חצה 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($S, k, \text{max_depth}, \text{min_samples_split}$):

For $i=1$ to k :

$S' = \text{Sample } \sqrt{d} \text{ features out of the original } d \text{ features in } S \text{ (keeping all samples)}$

$h_i = \text{ID3}(S', \text{max_depth}, \text{min_samples_split}, \text{criterion}=\text{"entropy"})$

Return $H(x) = \frac{1}{k} \sum_{i=1}^k h_i(x)$ *ממוצע של ה- h_i 'ים*

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

סמנו את כל התשובות הנכונות (השאלה אינה עוסקת במקרי קצה אלא במקרה הסביר). *כמה המונח של המסווג?*

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. ☒ *נורמליזציה*

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

(יש שאלה נוספת בעמוד הבא)

נתונה פונקציית מיפוי כלשהי $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^{10}$.

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

AVG

a. כן.

c. רק אם הפונקציה ϕ לא ליניארית.

$c = 1$ and d

e. $c \geq 1$ ואם $c < 1$

f. לא, כי חסר גורם רגולריזציה.

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

This image shows a single page from a notebook or ledger. The page is white with ten horizontal blue ruling lines spaced evenly apart. On the left side, there is a vertical margin line, also in blue. The top-left corner of the page is rounded. There are no markings, text, or drawings on the page.

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

A large rectangular box with rounded corners, containing 25 horizontal lines for writing. The box is empty, with the lines spaced evenly across the page.

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

A large rectangular box with rounded corners, containing 25 horizontal lines for writing. The box is empty, with the lines evenly spaced across its height.