הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ב – 10 בפברואר 2022

מרצה: ד"ר יונתן בלינקוב

# מבחן מסכם מועד א' – פיתרון חלקי

## הנחיות הבחינה:

- **משך הבחינה:** 3 שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- אין צורך במחשבון.
- מותר לכתוב בעט או בעיפרון, כל עוד הכתב קריא וברור.
- מותר לענות בעברית או באנגלית.
- יש לכתוב את התשובות **על גבי שאלון זה** בכתב יד קריא. תשובה בכתב יד לא קריא – לא תיבדק.
- במבחן 16 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.**

## מבנה הבחינה:

- **חלק א' [75 נק']:** 4 שאלות פתוחות.
- **חלק ב' [25 נק']:** 5 שאלות סגורות (אמריקאיות) [כל אחת 5 נק'].

# בהצלחה!

# חלק א' – שאלות פתוחות [75 נק']

## שאלה 1 [16 נק']

חוקרת מהטכניון עובדת על בעיית סיווג בינארי כלשהי. ברשותה dataset שבו $m = 150$ דוגמאות שונות (distinct).

החוקרת הריצה שלושה מודלים, ולכל מודל ביצעה hyperparameter tuning:

(a) מודל: kNN (נק' נחשבת שכנה של עצמה), היפרפרמטר: מספר השכנים $k$, טווח: 1 עד 97.
(b) מודל: עץ החלטה, היפרפרמטר: עומק מירבי, טווח: 1 עד 40.
(c) מודל: Kernel SVM, היפרפרמטר: חוזק הרגולריזציה $\lambda$, טווח: $10^{-3}$ עד $10^{7}$.

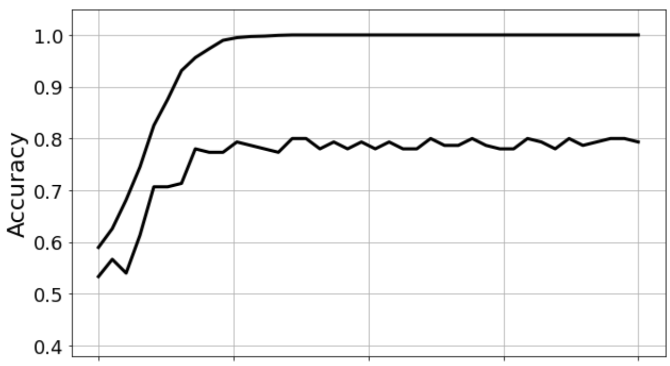
לכל מודל, היא ציירה גרף של דיוק האימון ודיוק ההכללה (בעזרת 5-fold cross validation) בציר $y$ כפונקציה של ערך ההיפרפרמטר בציר $x$ (הערכים גדלים משמאל לימין).

בעקבות תקלה, הכיתוב על ציר $x$ נמחק מכל הגרפים.

בנוסף, הגרפים נשמרו בטעות בשחור לבן, כך שלא ניתן להבדיל בקלות בין דיוק האימון לדיוק ההכללה.

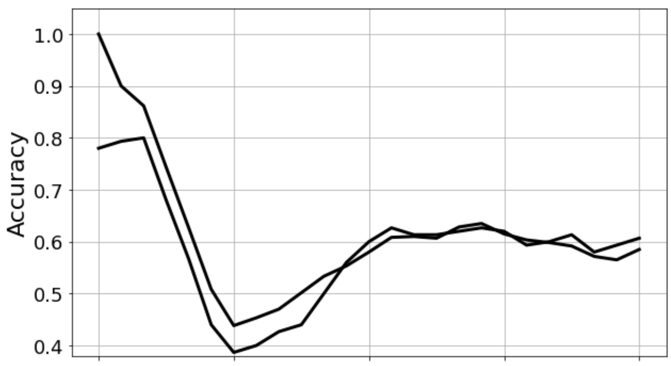
א. [8 נק'] במקום המתאים מתחת כל תרשים, כתבו את האות שמתאימה למודל ולהיפרפרמטר שיצרו אותו.

תרשים א'



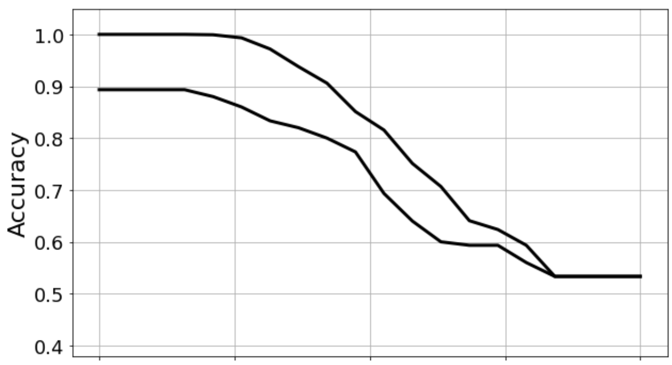

b

תרשים ב'




a

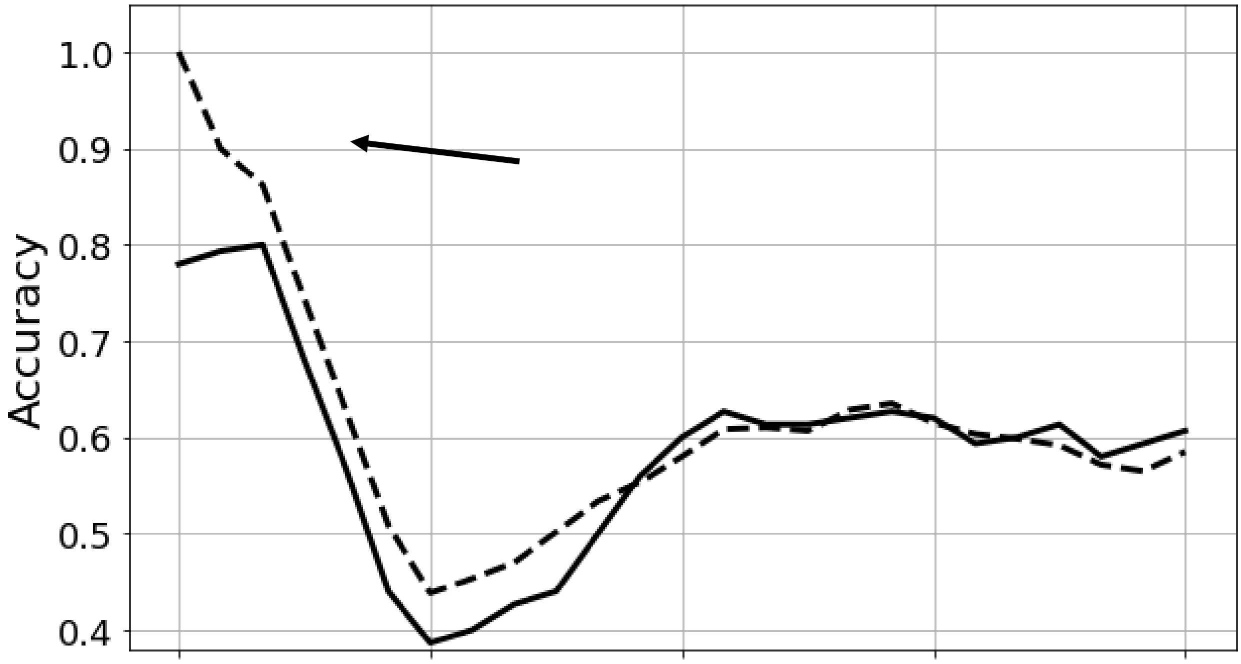
תרשים ג'




c

הערות בדיקה: loss האימון של SVM עם רגולריזציה אמור לעלות באופן מונוטוני ולכן התכוונו ש-SVM זה c.
עם זאת, זה לא גורר שה-error עצמו עולה (וה-accuracy יורד). ולכן בדיעבד קיבלנו גם תשובות שהחליפו בין a ל-c.

ב. [8 נק'] להלן תרשים ב' מוגדל.




הסתכלו על העקומה המְקוּוְקֶוֶת שבתרשים (מסומנת בחץ).

האם העקומה מתארת את דיוק האימון או את דיוק ההכללה?

הסבירו בקצרה. התבססו על התרשים ועל מאפייני המודל שיצר את עקומה זו (מבין שלושת המודלים).

דיוק האימון, כי kNN עם $k = 1$ חייב לתת דיוק אימון מושלם

(כאשר אין דוגמאות שונות ודוגמת אימון נחשבת שכנה של עצמה).

זה שבכלליות העקומה יותר גבוהה, לא אומר חד משמעית שזו עקומת האימון.

<u>הערות בדיקה</u>: זיהוי שמדובר בדיוק האימון זיכה ב-2 נק'. יתר הנקודות ניתנו להסבר.

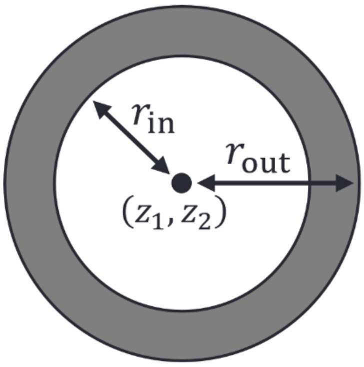
# שאלה 2 – PAC learning [21 נק']

תהי $\mathcal{H}$ מחלקת היפותזות של Bagels/donuts בדו-ממד:

$$\mathcal{H} = \{h_\theta : \mathbb{R}^2 \to \pm 1 \mid \theta = (z_1, z_2, r_{\text{out}}, r_{\text{in}}),\ r_{\text{out}} > r_{\text{in}} \geq 0\}$$

כאשר היפותזה בודדת מוגדרת באופן הבא:

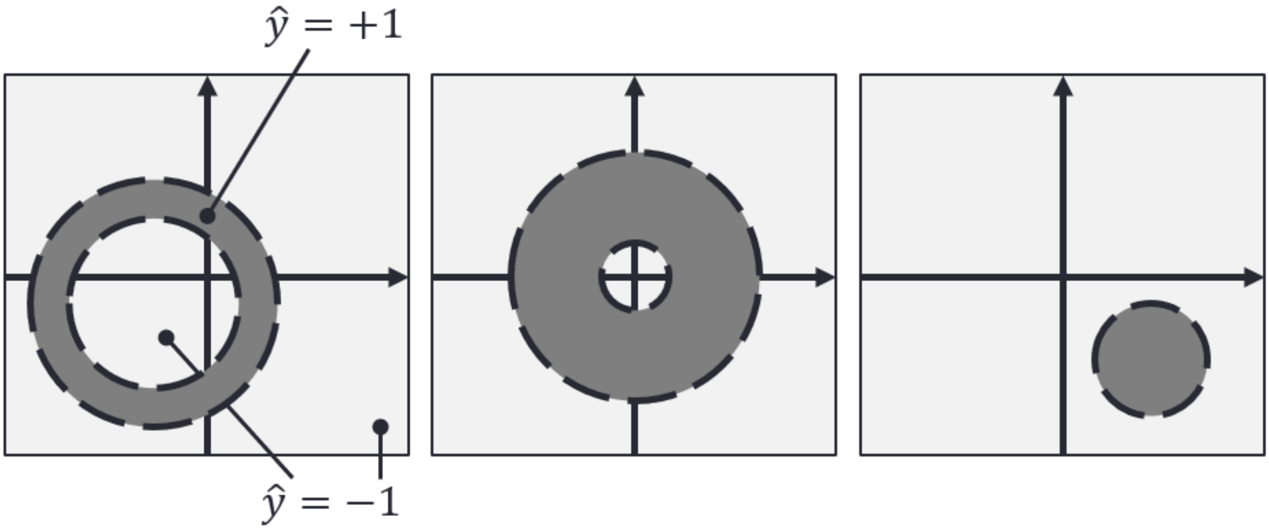
$$h_\theta(x) = \begin{cases} +1, & r_{\text{out}} \geq \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2} \geq r_{\text{in}} \\ -1, & \text{otherwise} \end{cases}$$




דגשים לגבי כל היפותזה $h_\theta \in \mathcal{H}$:

- המרכזים של המעגלים משותפים ולא בהכרח בראשית הצירים.
- הרדיוס של המעגל הפנימי יכול להיות אפס.
  השטח שבתוך ה-donut לא יכול להיות אפס.
- האזור בין שני המעגלים מסווג כחיובי, והאזורים האחרים כשליליים.
- מדובר אך ורק במעגלים ולא באליפסות.

דוגמה לשלוש היפותזות מתוך $\mathcal{H}$:



א. [3 נק'] להלן ההגדרה של "ניתוץ". השמטנו מההגדרה את הַכַּמָּתים.

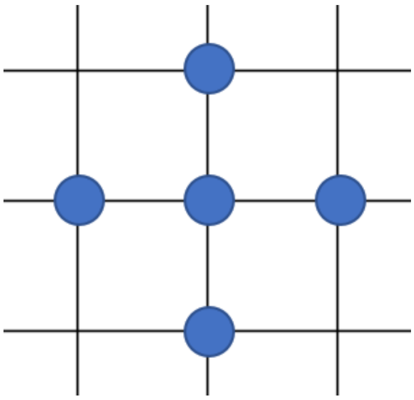
השלימו את שלושת הכמתים החסרים. בכל מקום כתבו האם חסר בהגדרה ∀ או ∃.

$$\mathcal{H} \text{ shatters } C \quad \Leftrightarrow \quad \forall y_1, \dots, y_{|C|} \in \mathcal{Y}:\ \exists h \in \mathcal{H}:\ \forall x_i \in C:\ h(x_i) = y_i$$

ב. [13 נק'] כתבו את החסם התחתון **<u>ההדוק ביותר</u>** שתוכלו למצוא לממד ה-VC: $\boxed{5}$ $\geq \mathrm{VCdim}(\mathcal{H})$.

הוכיחו את החסם התחתון שכתבתם (אין להוכיח שוויון). יש לכתוב הסבר מילולי <u>תמציתי</u> ולצרף תרשימים נדרשים.

דוגמאות לאוסף נקודות מנותץ:



<u>הערות בדיקה</u>: סטודנטים שכתבו $VC = 3$ קיבלו לכל היותר 5 נקודות.

סטודנטים שכתבו $VC = 4$ יכלו לקבל ניקוד מלא.

לסטודנטים שכתבו $VC = 5$ ירדו נקודות, אך בדיעבד מדובר בטעות שלנו ונתקן זאת בבדיקה מחודשת.

ג. [5 נק'] חוקרת וחוקר רוצים לאמן מודל סיווג בינארי.

<u>החוקרת</u> משתמשת במחלקת ההיפותזות $\mathcal{H}$ שהגדרנו.

<u>החוקר</u> משתמש במחלקת היפותזות של donuts שמרכזם בראשית הצירים, משמע:

$$\mathcal{H}' = \{h_\theta \mid \theta = (0, 0, r_{\text{out}}, r_{\text{in}}),\ r_{\text{out}} > r_{\text{in}} \geq 0\} \subset \mathcal{H}.$$

מי צפוי להזדקק לפחות דוגמאות בתהליך הלמידה ע"מ להבטיח (במונחי PAC) שגיאת הכללה $\epsilon = 0.1$? נמקו בקצרה.

תשובה תמציתית:

ה-VC של המחלקה החדשה הוא 2. החוקר יצטרך <u>פחות</u> דוגמאות לפי חסמי sample complexity.

## שאלה 3 – רגרסיה ליניארית ו-Generative models [21 נק']

נתון דאטה $S=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m$ שהגיע ממודל ליניארי $y_i=\boldsymbol{w}^\top\boldsymbol{x}_i+\varepsilon_i$ עם רעש אקראי מפילוג i.i.d. נורמלי: $\varepsilon_i\sim\mathcal{N}(0,\ 1)$.

<u>שימו לב</u>: הדוגמאות $\boldsymbol{x}_i\in\mathbb{R}^d$ והתיוגים $y_i\in\mathbb{R}$ נתונים. וֶקטור המשקלים $\boldsymbol{w}\in\mathbb{R}^d$ לא ידוע ואותו אנו רוצים ללמוד.

<u>תזכורת</u>: הוכחנו שתחת הנחות אלה ה-likelihood שווה ל: $\Pr\left(\left\{\boldsymbol{x}_i,y_i\right\}_i\middle|\boldsymbol{w}\right)=\prod_{i=1}^m\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{w}^\top\boldsymbol{x}_i-y_i\right)^2\right\}$

א. [5 נק'] הוכיחו שתחת הנחות השאלה, בעיית ה-LS ללא רגולריזציה, משמע $\underset{\boldsymbol{w}}{\operatorname{argmin}}\frac{1}{m}\sum_{i=1}^m(\boldsymbol{w}^\top\boldsymbol{x}-y_i)^2$, שקולה לבעיית ה-MLE הבאה: $\underset{\boldsymbol{w}}{\operatorname{argmax}}\,L(\boldsymbol{w}|\{\boldsymbol{x}_i,y_i\}_i)$. הוכחה כמו בתרגול.

כעת, נניח <u>בנוסף</u> **שווקטור המשקלים** הלא ידוע $\boldsymbol{w}$ הגיע מהתפלגות <u>לפלאס</u>. משמע, כל משקל נדגם <u>i.i.d</u> באופן הבא: $\forall k=1,\dots,d\colon w_k\sim\text{Laplace}(0,b)$, עבור $b>0$ נתון (משותף לכל המשקלים).
שימו לב: עדיין מניחים שהרעש $\epsilon_i$ מתפלג גאוסיאנית ככתוב בתחילת השאלה.

פונקציית הצפיפות של התפלגות לפלאס הינה: $Z\sim\text{Laplace}(\mu,b)\Longrightarrow f(z)=\frac{1}{2b}\exp\left\{-\frac{1}{b}|z-\mu|\right\}$.

ב. [11 נק'] הוכיחו שתחת <u>כלל</u> ההנחות, בעיית LS עם רגולריזציית $\ell^1$, משמע $\underset{\boldsymbol{w}}{\operatorname{argmin}}\frac{1}{m}\sum_{i=1}^m(\boldsymbol{w}^\top\boldsymbol{x}-y_i)^2+\lambda\|\boldsymbol{w}\|_1$, שקולה לבעיית MAP תחת ה-prior על המשקלים, משמע $\underset{\boldsymbol{w}}{\operatorname{argmax}}\Pr\left(\boldsymbol{w}\,\middle|\,\left\{\boldsymbol{x}_i,y_i\right\}_i,\ \mu=0,\ b\right)$.

Like we did in class for other distributions:

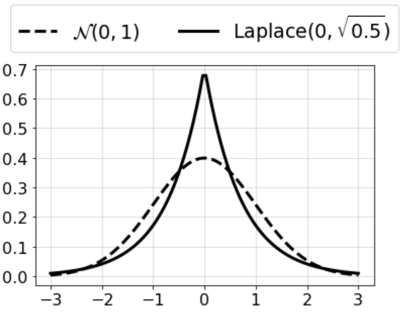
$$\widehat{\boldsymbol{w}}_{MAP}\triangleq\underset{\boldsymbol{w}}{\operatorname{argmax}}\,p(\boldsymbol{w}|\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m,\mu=0,b)=\underset{\boldsymbol{w}}{\operatorname{argmax}}[p(\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m|\boldsymbol{w})\cdot p(\boldsymbol{w}|\mu=0,b)]$$

$$=\underset{\boldsymbol{w}}{\operatorname{argmax}}\ln[p(\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m|\boldsymbol{w})\cdot p(\boldsymbol{w}|\mu=0,b)]$$

$$=\underset{\boldsymbol{w}}{\operatorname{argmax}}\ln\left[(2\pi)^{-\frac{m}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^m(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle-y_i)^2\right\}\right]+\ln\left[(2b)^{-d}\exp\left\{-\frac{1}{b}\|\boldsymbol{w}\|_1\right\}\right]$$

Get rid of the additive constant (w.r.t to $\boldsymbol{w}$) and get:

$$=\underset{\boldsymbol{w}}{\operatorname{argmax}}-\frac{1}{2}\sum_{i=1}^m(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle-y_i)^2-\frac{1}{b}\|\boldsymbol{w}\|_1=\underset{\boldsymbol{w}}{\operatorname{argmin}}\left(\frac{1}{m}\sum_{i=1}^m(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle-y_i)^2+\frac{2}{mb}\|\boldsymbol{w}\|_1\right)$$




השונות של התפלגות לפלאס נתונה ע"י $\text{Var}[w_k]=2b^2$.

התרשים משווה בין התפלגות לפלאס להתפלגות נורמלית שהשונות שלהן היא 1.

<u>תזכורת</u>: הוכחנו בתרגול שאם מניחים $w_k\sim\mathcal{N}(0,1)$, בעיית ה-MAP שקולה לבעיית LS עם רגולריזציית $\ell^2$.

ג. [5 נק'] מתוך הסתכלות בתרשים, מתוך התזכורת ומתוך מה שהוכחתם בסעיף הקודם, הסבירו בקצרה ובאופן אינטואיטיבי (לא פורמלי) הבדל שלמדנו בין אופי הפיתרונות שמתקבלים ע"י רגולריזציית $\ell^2$ לאלה המתקבלים ע"י רגולריזציית $\ell^1$.

LASSO מביא לפתרונות יותר דלילים. ניתן לראות זאת בתרשים כי ההתפלגות יותר צפופה סביב האפס

## שאלה 4 – Kernel SVM [17 נק']

עבור פרמטר נתון $\gamma > 0$, נגדיר את ה-Gaussian kernel לקלט חד-ממדי באופן הבא:

$$K: \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad K(a, b) = \exp(-\gamma(a-b)^2)$$

א. [12 נק'] הציעו פונקציית מיפוי $\phi: \mathbb{R} \to \mathbb{R}^p$ והוכיחו בעזרתה שהפונקציה $K$ מהווה קרנל חוקי (בחד ממד).

שימו לב: עליכם לבחור $p \in \mathbb{N} \cup \{\infty\}$ מתאים, סופי או אינסופי.

תזכורת: פירוק טור טיילור של $e^x$ לכל $x$ נתון ע"י: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.

תשובה:

$$exp(-\gamma(a-b)^2) = exp\left(-\gamma(a^2 - 2ab + b^2)\right) = exp(-\gamma a^2)\, exp(-\gamma b^2)\, exp(2\gamma ab)$$

$$= exp(-\gamma a^2)\, exp(-\gamma b^2) \sum_{n=0}^{\infty} \frac{(2\gamma ab)^n}{n!} = exp(-\gamma a^2)\, exp(-\gamma b^2) \sum_{n=0}^{\infty} \frac{(2\gamma)^n a^n b^n}{n!}$$

And we propose: $\phi: \mathbb{R} \to \mathbb{R}^\infty$, $\phi_n(a) = exp(-\gamma a^2) \frac{(2\gamma)^{\frac{n}{2}} a^n}{\sqrt{n!}}$

ב. [5 נק'] נתון dataset עם $m = 1000$ דוגמאות חד-ממדיות. נרצה לפתור את הבעיה עם ה-Gaussian kernel שהגדרנו. מבחינת יעילות, האם עדיף לפתור את ה-primal problem עם ה-feature mapping שמצאתם, או שעדיף לפתור את ה-dual problem עם פונקציית ה-kernel שהוגדרה? ענו והסבירו בקצרה.

תשובה:

לא ניתן לפתור את הבעיה ב-primal כי מרחב האופטימיזציה המתאים הוא אינסוף ממדי.

# חלק ב' – שאלות אמריקאיות [25 נק']

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

א. [5 נק'] סמנו את כל התשובות הנכונות ביחס לאלגוריתמי One vs. All (1vA) ו-One vs. One (1v1)
(הניחו שיש 10 מחלקות ומעלה).

a. *ל-1vA סיבוכיות מקום נמוכה יותר מזו של 1v1 בזמן האימון.*

b. *ל-1vA סיבוכיות מקום נמוכה יותר מזו של 1v1 בזמן המבחן (לאחר שהאימון הושלם).*

c. רק אחד משני האלגוריתמים ניתן למיקבול (parallelization).

d. 1v1 נוטה יותר ליצור בעיות לא מאוזנות (imbalanced).

הערות בדיקה: עבור בעיה עם $K$ מחלקות, 1vA מאמן $K$ מסווגים ואילו 1v1 מאמן $\binom{K}{2}$ מסווגים.

טעות אחת (למשל, לא לסמן את a) הובילה להפחתה של 2 נקודות. שתי טעויות נוקדו בהתאם לחומרתן.

ב. [5 נק'] סמנו את כל הטענות שמשלימות בצורה הגיונית את הטענה הבאה.
באופן כללי, כְּכֹל שה-complexity של מחלקת היפותזות עולה:

a. ה-bias עולה.

b. *ה-variance עולה.*

c. צריך פחות דאטה על מנת להכליל כראוי.

d. *יש יותר נטייה ל-overfitting.*

e. תהליך האימון של מסווג בודד דורש זמן רב יותר.

הערות בדיקה: תשובה e לא נכונה באופן כללי, אבל סימון שלה גרר הפחתה של נקודה אחת בלבד.

ג. [5 נק'] נגדיר את פונקציית ה-squared hinge loss: $\mathcal{L}(z) = (\max\{0,\ 1 - z\})^2$.
סמנו את כל הטענות הנכונות ביחס לפונקציה זו.

a. *הפונקציה קמורה ביחס ל-$z$.*

b. הנגזרת של הפונקציה היא $\frac{\partial}{\partial z}\mathcal{L} = 2 - 2z$.

c. *הפונקציה חוסמת מלמעלה את ה-0-1 loss בכל מקום.*

d. הפונקציה חוסמת מלמעלה את ה-hinge loss בכל מקום.

e. *עבור בעיות סיווג תחת מודל ליניארי, משמע $z = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$, הפונק' מעוֹדדת margin מהמפריד.*

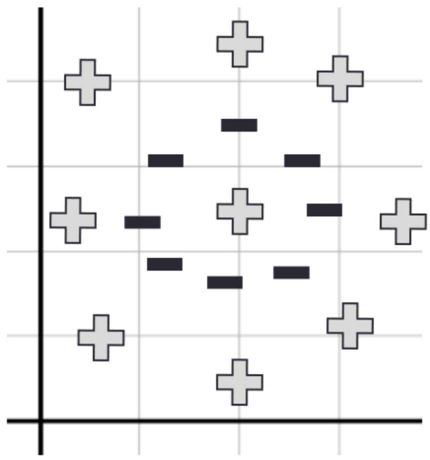
הערות בדיקה:

הפונקציה "מענישה" עבור $y_i \boldsymbol{w}^\top \boldsymbol{x}_i < 1$ והמשמעות היא שהיא מעודדת margin (בדיוק כמו hinge loss רגיל).

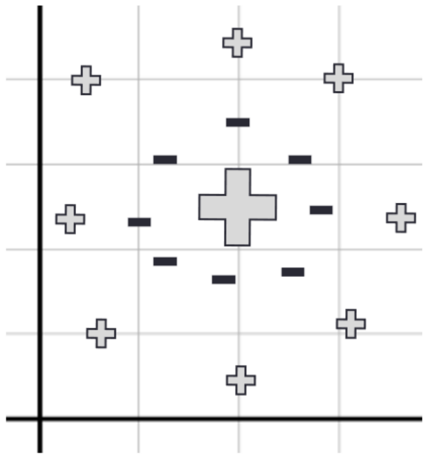
טעות אחת (למשל, לא לסמן את e) הובילה להפחתה של 2 נקודות. שתי טעויות הובילו לציון 0 בשאלה.

ד. [5 נק'] נתון דאטה עם תיוגים בינאריים ("+" או "-"). מריצים AdaBoost עם מסווג בסיס לא ידוע.
גדלי הצורות בתרשימים מסמלים את ההסתברויות שהאלגוריתם מקצה (הסתברות גבוהה = צורה גדולה).
מריצים את האלגוריתם איטרציה אחת ומקבלים את התרשימים הבאים:

**ההתפלגות האחידה ההתחלתית**



**ההתפלגות לאחר איטרציה אחת**



(הנק' האמצעית גדלה והיתר קטנו)

איזה סוג של מסווג בסיס יכול להסביר את התרשים השמאלי שהתקבל? סמנו את התשובה הנכונה.

a. עץ החלטה בעומק 1 (decision stump). משמע, שורש ושני עלים.

b. עץ החלטה בעומק 2. משמע, שורש, רמת ביניים ועד ארבעה עלים.

c. מסווג שאומר על כל המרחב "שקר" או "אמת".

*d. SVM עם קרנל פולינומיאלי ממעלה 2.*

e. כל התשובות הקודמות לא נכונות.

ה. [5 נק'] היזכרו בפונקציית ה-Softmax שמשמשת כשכבה האחרונה של רשת נוירונים לסיווג ל-$K$ מחלקות:

$$\mathrm{softmax}(f_1(x),\dots,f_K(x);\ \beta\ ) = \left[\frac{\exp\{\beta f_1(x)\}}{\sum_{i\in[K]}\exp\{\beta f_i(x)\}}\ ,\ \ \dots\ \ ,\ \frac{\exp\{\beta f_K(x)\}}{\sum_{i\in[K]}\exp\{\beta f_i(x)\}}\right]^{\top}$$

בסעיף זה אנו לא מתייחסים כלל לאפשרות ש-$\beta = 0$ ומניחים של-$\beta$ אותו סימן בזמן האימון ובזמן המבחן.

סמנו את כל הטענות הנכונות ביחס לפונקציה זו.

a. בזמן מבחן (לאחר האימון), כאשר $\beta \to \infty$, התפלגות הפלט הולכת להתפלגות אחידה.

*b. כאשר משנים את $\beta$ לאחר האימון בזמן המבחן, כל עוד $\beta$ שומר על הסימן, אין לו השפעה על הדיוק של הרשת.*

c. בזמן אימון, כל עוד הפרמטר $\beta$ חיובי, אין לו השפעה על מהלך האימון.

d. בזמן אימון, אם הפרמטר $\beta$ שלילי, לא ניתן ללמוד את הרשת בעזרת שיטות gradient.

הערות בדיקה: סימון של תשובה שגויה בנוסף ל-b, הובילה להפחתה של שתי נקודות.