הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ד – 3 בספטמבר 2024

מרצה: ד"ר ניר רוזנפלד

פתרון

<u>מבחן מסכם מועד א'</u>

הנחיות הבחינה:

- **משך הבחינה:** שלוש שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- מחשבון: מותר.
- כלי כתיבה: עט <u>בלבד</u>.
- יש לכתוב את התשובות **על גבי שאלון זה**.
- מותר לענות בעברית או באנגלית.
- הוכחות והפרכות צריכות להיות פורמליות.
- קריאות:
  - סימונים לא ברורים בשאלות רב-ברירה ו/או תשובות מילוליות בכתב יד לא קריא יובילו לפסילת התשובה.
  - לא יתקבלו ערעורים בנושא.
- במבחן 19 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.**
- **לזכאים להערכה חלופית מתאפשרת בחירה בין שאלות 3 ו-4.**
- **זכרו: Less is more**. אל תכתבו פרטים מיותרים.

בהצלחה!

שאלה 1: SVM, Imbalanced sampling [35 נק']

בבעיה זאת נעסוק בבעיית סיווג בינארי $\mathcal{Y} = \{-1,1\}$, עם מרחב דוגמאות $\mathcal{X} = \mathbb{R}^d$.

א. [15 נק'] בסעיף זה $d = 2$. דגמו באקראי $S \subset \mathcal{X}$.
מימין נתונים 4 איורים המראים decision boundaries שונים. באזורים אפורים המודל מנבא $+1$.
משמאל נתונות 4 גרסאות שונות של בעיות SVM. בנוסף נתונה פונקציה למיפוי פיצ׳רים פולינומיאלית ממימד גבוה ϕ.

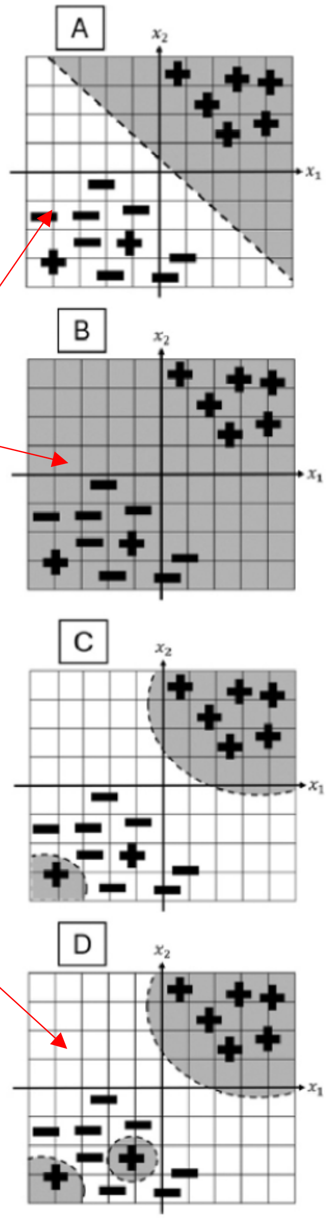
1) ע״י מתיחת קו, לכל בעיית SVM התאימו איור אחד שמייצג את ה-decision boundary של המודל שיתקבל מבעיה זו. אם בעיית ה-SVM לא יכולה להגיע לפתרון, סמנו עליה [X].

<u>תזכורת</u>: כלל ההחלטה עבור מפריד ליניארי ללא מיפוי ϕ - $h(\mathbf{x}) = sign(\mathbf{w}^\top \mathbf{x} + b)$.

עם מיפוי ϕ - $h(\mathbf{x}) = sign(\mathbf{w}^\top \phi(\mathbf{x}) + b)$.

במהלך המבחן הובהר כי $sign(0) = +1$.

I) $\underset{\mathbf{w},b}{\operatorname{argmin}} \ \|\mathbf{w}\|_2^2 + \|b\|$
II) $\underset{\mathbf{w},b}{\operatorname{argmin}} \ \|\mathbf{w}\|_2^2$ such that: $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \forall i$
III) $\underset{\mathbf{w},b}{\operatorname{argmin}} \ \|\mathbf{w}\|_2^2$ such that: $y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1, \forall i$
IV) $\underset{\mathbf{w},b}{\operatorname{argmin}} \ \|\mathbf{w}\|_2^2 + \sum_i \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}$



2) נמקו בקצרה את בחירתכם.

פתרון:

I – הפתרון שיביא את ה-*objective* למינימום הוא $\mathbf{W} = \mathbf{0}, b = 0$, מה שיוצר את הפתרון $sign(0) = +1$ לכל $\mathbf{x}_i$.

II – זוהי בעיית *hard SVM*, אך מכיוון שהדאטא לא פריד לינארית לא ניתן להגיע לפתרון.

III – זהו *hard SVM* עם מיפוי פיצ'רים ϕ. כלומר הוא חייב לסווג את כל הנקודות נכון, והדבר מתאפשר שכן מיפוי הפיצ'רים ממימד גבוה.

IV – זהו *soft SVM*, כלומר הוא יגיע לפתרון לינארי המסוגל לטעות על חלק מהדוגמאות.

מבין ארבעת בעיות ה- SVM, עבור מי הביטוי $\mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right]$ הוא הגדול ביותר? נמקו בקצרה.

הערה: התוחלת כאן היא ביחס למשתנים המקריים S, x.

תזכורת: הפונקציה $\bar{h}(x)$ היא המסווג הממוצע של מחלקת ההיפותזות המוגדרת ע״י בעיית ה-SVM.

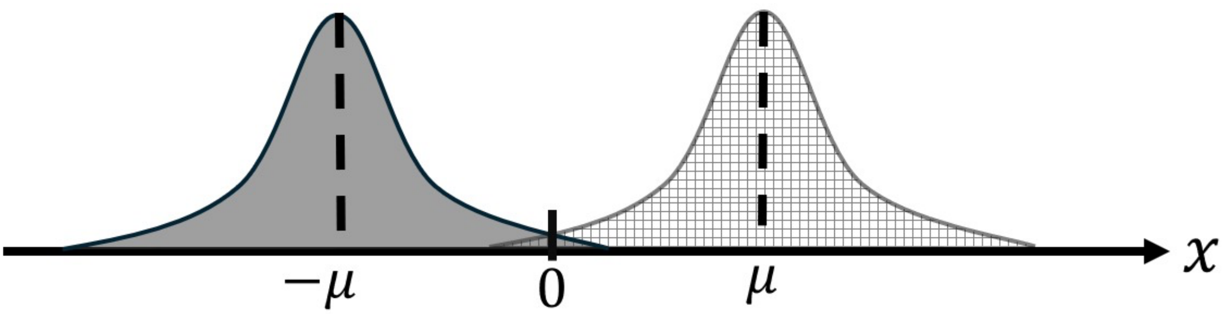
פתרון: זהו ה-*variance* שלמדנו בתרגול של *Model selection*. למדנו כי ביטוי זה גדל ככל שמחלקת המודלים עשירה יותר. במקרה זה, בעיית ה-*SVM* בעלת המחלקה העשירה ביותר היא בעיה III, שכן מדובר בבעיה היחידה עם מיפוי פיצ'רים. כמו כן מדובר בבעיית ה-*hard SVM*, כלומר היא חייבת להגיע להתאמה מושלמת על הדאטה.

ב. [20 נק׳] בסעיפים הבאים ננתח תרחיש של imbalanced sampling, תרחיש שבו התפלגות התיוגים בקבוצת מדגם S היא לא מאוזנת.

ננתח את המקרה שבו $d = 1$. ידוע כי הפיצ׳ר x מתפלג כמו משתנה אקראי גאוסי, כאשר תוחלת הגאוסיאן תלויה בערך של y. השונות זהה בשני המקרים. יהי $\mu > 0$, אזי:

- $\mathbb{P}(x|y = 1) = N(\mu, \sigma^2)$.
- $\mathbb{P}(x|y = -1) = N(-\mu, \sigma^2)$.
- $\mathbb{P}(y = 1) = \mathbb{P}(y = -1) = 1/2$.

1) להלן שרטוט של התפלגות המשותפת (עם צביעה שונה להתפלגויות המותנות השונות):



נניח כי יש לנו גישה להתפלגות עצמה, כלומר אנחנו יודעים את μ, σ^2. מהו המפריד מסוג threshold (לינארי) אשר ישיג את שגיאת ההכללה המינימאלית על פילוג זה.

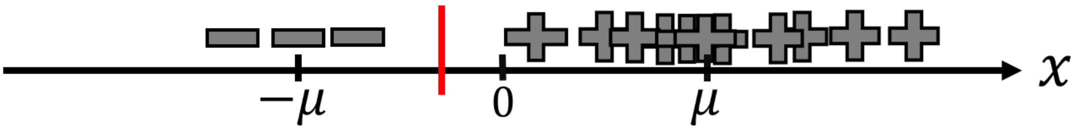
כתבו את כלל ההחלטה שלו (כלומר כלל הפרדיקציה $\hat{y} = h(x)$) באופן מפורש, ונמקו בקצרה.

פתרון: משיקולי סימטריה המפריד בעל שגיאת ההכללה המינימאלית יעבור בדיוק במרכז בין שני הגאוסיאניים, כלומר ב-0. המודל הינו $h(x) = sign(x)$.

נניח כעת כי יש לנו גישה רק לדגימה מההתפלגות (כלומר אנחנו לא יודעים את $(\mu,\sigma^2$. יהי $S=\{(x_i,y_i)\}_{i=1}^{n}$ מדגם בגודל n אשר כל דגימה (x_i,y_i) נדגמה $i.i.d$ מההתפלגות המשותפת $\mathbb{P}(x,y)$. נסמן ב-n_+ את מספר הדוגמאות החיוביות במדגם, וב-n_- את מספר הדוגמאות השליליות, כך ש- $n=n_+ + n_-$.

נגיד שהמדגם "לא מאוזן" כאשר n_+ גדול משמעותית מ-n_- (או להפך).

2) להלן איור של מדגם מסוים S שאינו מאוזן. נניח שלמדנו בעזרת מסווג threshold עם אלגוריתם SVM. סמנו על האיור את המקום בו עובר המפריד הנלמד. נמקו את בחירתכם, והסבירו את ההשלכות של מדגם לא מאוזן על תוצאות הלמידה.



<u>פתרון</u>: על מנת למקסם את ה-*margin*, ה-*SVM* יעביר את המפריד במרחקים שווים בין הנקודה החיובית השמאלית ביותר לנקודה השלילית הימנית ביותר (אלו ה-*support vectors*).
מכיוון שהמדגם לא מאוזן אנו רואים יותר דגימות מההתפלגות של $y=+1$, מה שמגדיל את הסיכוי לראות ערכים רחוקים יותר מתוחלת הגאוסיאן. בשל אופי ה-*SVM* דבר זה גורם למפריד לסטות שמאלה מהערך האופטימלי של 0; לכן נקבל מודל עם שגיאת הכללה לא אופטימלית ועם הטיה (*bias*) נגד הדגימות של $y=-1$ (כלומר הוא יטעה על הרבה מהן).

<u>טעויות נפוצות</u>: דיון חסר בנוגע להשלכות של המדגם הלא מאוזן על המודל. לדוגמה:
1. לא לציין שזה יגרום למודל הנלמד להיות עם שגיאת הכללה לא אופטימלית.
2. לא לציין שלמודל הנלמד תהיה הטיה נגד הדגימות של $y=-1$.

3) עבור התפלגות נתונה ומדגם בגודל n ממנה, נגדיר את הגדלים הבאים:
- $M(n)=\mathbb{E}[\max(x_1,\dots,x_n)]$.
- $m(n)=\mathbb{E}[\min(x_1,\dots,x_n)]$.

כאשר מדובר בהתפלגות גאוסיאנית עם תוחלת μ (ושונות קבועה σ^2), נסמן ב-$M(n;\mu)$ ו-$m(n;\mu)$ את הגדלים המתאימים.

נתון שמתקיים בקירוב: $M(n;0)\cong\sigma\sqrt{2\ln(n)}$, $m(n;0)\cong-\sigma\sqrt{2\ln(n)}$.

השלימו:

$M(n;-\mu)\cong\sigma\sqrt{2\ln(n)}-\mu$	$M(n;\mu)\cong\sigma\sqrt{2\ln(n)}+\mu$
$m(n;-\mu)\cong-\sigma\sqrt{2\ln(n)}-\mu$	$m(n;\mu)\cong-\sigma\sqrt{2\ln(n)}+\mu$

4) נתון שדגמתם מדגם לא מאוזן אך פריד. שיטה נפוצה להתמודדות עם בעיה זו נקראת subsampling.
במסגרת שיטה זו, מוציאים באקראי מ-S דוגמאות **בעלות התיוג הנפוץ ביותר**, עד שמקבלים קבוצת מדגם מאוזנת , כלומר עד ש-$n_+ = n_-$.

השתמשו בסעיפים הקודמים, ובפרט בסעיף (3), על מנת להסביר כיצד subsampling יכול לסייע לאימון SVM בתרחיש של שאלה זו.
הדרכה: שימו לב כי הגדלים $M(n), m(n)$ תלויים ב- n. בנוסף זכרו מה התכונה של המפריד הלינארי שמחזיר SVM.

פתרון: כפי שראינו בסעיפים קודמים, ה-*SVM* יעביר את המפריד במרחקים שווים בין הנקודה החיובית השמאלית ביותר לנקודה השלילית הימנית ביותר (אלו ה-*support vectors*). במקרה שלנו מדובר ב-$m(n_+;\mu)$ וב-$M(n_-;-\mu)$ בהתאמה. לאחר ה-*subsampling* מתקיים $n_+ = n_-$, כלומר ניתן לחשב מפורשות היכן בתוחלת יעבור המפריד:

$$\frac{m(n_+;\mu) + M(n_-;-\mu)}{2} = \frac{-\sigma\sqrt{2\,ln(n_+)} + \mu + \sigma\sqrt{2\,ln(n_-)} - \mu}{2} = 0$$

כלומר בזכות ה-*subsampling* קיבלנו את המפריד עם שגיאת ההכללה האופטימלית.

טעויות נפוצות:
1. דיון כללי על איך איזון הדאטה עוזר, ללא שימוש מפורש בנוסחאות של $(n_+;\mu)$ ו-$M(n_-;-\mu)$ מסעיף קודם.
2. חוסר התייחסות למצב של $n_+ = n_-$
3. נימוק המשתמש ב-$|m(n_+;\mu) - M(n_-;-\mu)|$ או בטענה כללית שה-*support vectors* "מתרחקים" זה מזה כתוצאה מה-*subsampling*. דבר זה כשלעצמו אינו אומר שנקבל מודל עם שגיאת הכללה טובה יותר.

שאלה 2: VC-dimension [25 נק']

א. [2 נק'] להלן ההגדרה של "ניתוץ". השמטנו מההגדרה את הַכַּמָּתים.

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

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

בשאלה זו נתון מרחב דוגמאות $\mathcal{X}$ כלשהו ומרחב תיוגים $\mathcal{Y} = \{-1, +1\}$. אם נדרשת הוכחה, הוכיחו באופן פורמלי. אם נדרשת הפרכה, תנו דוגמה נגדית מנומקת היטב והוכיחו כי היא אכן מפריכה את הטענה.

ב. [5 נק'] הוכיחו/הפריכו.

יהיו שתי מחלקות $\mathcal{H}_1, \mathcal{H}_2$. אזי

$$VCdim(\mathcal{H}_1 \cup \mathcal{H}_2) \geq \max\{VCdim(\mathcal{H}_1), VCdim(\mathcal{H}_2)\}$$

פתרון: נניח בה"כ כי $VCdim(\mathcal{H}_1) \geq VCdim(\mathcal{H}_2)$. מהגדרת $VCdim$ קיימת קבוצה C בגודל $VCdim(\mathcal{H}_1)$ ש-$\mathcal{H}_1$ מנתצת.

נבחר בדיוק את אותה ה- C. יהי $y_1, \dots, y_{|C|}$ תיוג כלשהו שלה. קיימת $h \in \mathcal{H}_1$ כך שלכל $i = 1, \dots, |C|$ מתקיים $h(x_i) = y_i$.

כמובן שמתקיים $h \in \mathcal{H}_1 \cup \mathcal{H}_2$, ולכן לכל תיוג של C קיימת היפותזה מתאימה ב- $\mathcal{H}_1 \cup \mathcal{H}_2$ שצודקת על תיוג זה.

טעויות:

1. שימוש בתכונת המונוטוניות ללא הוכחה. שימו לב, זו שאלה ממבחן עבר ולא מההרצאות או התרגולים.
2. שימוש בניסוחים לא נכונים. למשל: זה לא נכון להגיד ש-h (היפותזה) מנתצת את C.
3. חלקכם ניסיתם לסתור טענה זו. אי אפשר.

ג. [6 נק'] הוכיחו.

נתונה מחלקה $\mathcal{H}$ סופית. אזי

$$VCdim(|H|) \leq \log_2 |H|.$$

פתרון: נסמן $VCdim(\mathcal{H}) = d$. אם כך, אז קיימת קבוצה C בגודל d ש-$\mathcal{H}$ מנתצת אותה. לכן לכל תיוג של קבוצה C קיימת היפותזה ב- $\mathcal{H}$ כך שההיפותזה צודקת על תיוג זה. מכיוון שהתיוגים שונים, מזה נגזר כי ההיפותזות לכל תיוג שונות ביניהן (כי הן פונקציות). סה"כ נקבל,

$$|\mathcal{H}| \geq 2^d \Rightarrow \log_2|\mathcal{H}| \geq d \Rightarrow \log_2|\mathcal{H}| \geq VCdim(\mathcal{H})$$

טעות נפוצה: הרבה מכם ניסיתם להוכיח טענה זו מתוך הנחה בשלילה. הטענה בשלילה כאן היא:

$$VCdim(\mathcal{H}) > \log_2|\mathcal{H}|$$

ואז הגעתם למסקנה כי בהכרח

$$VCdim(\mathcal{H}) \geq \log_2|\mathcal{H}| + 1$$

אך זה לא נכון, הרי אף אחד לא הבטיח לכם כי $\log_2 |\mathcal{H}| \in \mathbb{Z}$. הדבר הנכון לכתוב היה:

$$VCdim(\mathcal{H}) \geq \lceil \log_2 |\mathcal{H}| \rceil$$

ולהמשיך לעבוד משם.

ד. [12 נק'] הוכיחו/הפריכו.

יהיו $\mathcal{H}_1, \mathcal{H}_2$ מחלקות ונתונה $\emptyset \neq C \subset \mathcal{X}$. נניח

1) $\mathcal{H}_1$ מנתצת את C.

2) $\mathcal{H}_2$ אינה מנתצת את C.

אזי בהכרח

$$VCdim(\mathcal{H}_1) > VCdim(\mathcal{H}_2)$$

פתרון: נבחר $\mathcal{X} = \mathbb{R}$, ונגדיר את הפונקציות הבאות

$$h_0(x) = -1, \qquad h_1(x) = \begin{cases} +1, & x = 1 \\ -1, & o.w. \end{cases}, \qquad h_2(x) = \begin{cases} +1, & x = 2 \\ -1, & o.w. \end{cases}$$

נגדיר

$$\mathcal{H}_1 = \{h_0, h_1\}, \quad \mathcal{H}_2 = \{h_0, h_2\}$$

ונבחר $C = \{1\}$.

נבחין כי $\mathcal{H}_1$ מנתצת את C. עבור תיוג $+1$ מתקיים $h_1(1) = +1$ ועבור תיוג -1 מתקיים $h_0(1) = -1$.

נבחין כי $\mathcal{H}_2$ אינה מנתצת את C. עבור תיוג $+1$ מתקיים $h_0(1) = h_2(1) = -1$.

נבחין גם כי $\mathcal{H}_2$ מנתצת את הקבוצה $\{2\}$. עבור תיוג $+1$ מתקיים $h_2(2) = +1$ ועבור תיוג -1 מתקיים $h_0(2) = -1$.

לכן, $VCdim(\mathcal{H}_1) \geq 1, VCdim(\mathcal{H}_2) \geq 1$. מסעיף קודם נקבל,

$$VCdim(\mathcal{H}_1) \leq \log_2 2 = 1, \qquad VCdim(\mathcal{H}_2) \leq \log_2 2 = 1$$

וזו הסתירה המתבקשת.

טעויות נפוצות: קיימות אינספור מחלקות פשוטות שסותרות סעיף זה.

בשאלה זו נבחנתם על כמה דברים – 1) הגדרה ברורה עבור $\mathcal{H}_1, \mathcal{H}_2$. 2) בחירת C מתאימה.

3) להראות כי $\mathcal{H}_1$ מנתצת את C אך $\mathcal{H}_2$ לא מנתצת את C. 4) להראות כי $VCdim(\mathcal{H}_1) \leq VCdim(\mathcal{H}_2)$.

1. חלקכם לא הגדרתם כראוי את $\mathcal{H}_1, \mathcal{H}_2$. לא היה ברור האם מדובר במחלקה עם היפותזה בודדת, או מחלקה עם אינסוף היפותזות. כמו כן, הגדרת היפותזות מורכבות מדי (כמו למשל מפרידים לינאריים ב- $\mathbb{R}^{200}$) שגררו אחריהן הוכחות לא נכונות בהקשר לניתוץ.
2. חלקכם לא הגדרתם במקום ברור את C.
3. חלקכם לא הוכחתם כי $\mathcal{H}_1$ מנתצת את ה-C שבחרתם ו-$\mathcal{H}_2$ לא מנתצת אותה (או שלא עברתם על כל המקרים). בדר"כ זה נבע מבחירת מחלקות מורכבות מדי. הוכחה לא נכונה של ניתוץ.
4. חלקכם נשענתם על הוכחות $VCdim$ שלא ראינו בתרגולים ובהרצאות. שימו לב: הוכחות שהיו במבחני עבר לא נחשבות.

שאלה 3: Perceptron [20 נק']

☐ לזכאים להערכה חלופית בלבד (כפי שהוגדרו באתר הקורס): סמנו את התיבה הזו אם ברצונכם **לדלג** על שאלה זו. המשקל של יתר השאלות יתפזר באופן יחסי על פני 100 נקודות. ניתן לדלג רק על שאלה אחת מתוך שאלות 3,4.

בשאלה זו נניח $\mathcal{Y}=\{+1,-1\}$ ומרחב דוגמאות $\mathcal{X}=\mathbb{R}^d$. לפניכם אלגוריתם הפרספטרון כפי שהוא הוצג בתרגול.

```
input: training set S = {(x_1, y_1), ..., (x_m, y_m)}, step size η = 1

w = 0_d
while did not separate the training set:
      for i = 1 to m:
            ŷ_i = sign(wᵀx_i)
            if y_i ≠ ŷ_i:
                  w = w + y_i x_i
return w
```

לפניכם **משפט** התכנסות הפרספטרון:

תהי קבוצת אימון $S=\{(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_m,y_m)\}$. נניח כי קיימים וקטור משקולות w_* כך ש- $\|\mathbf{w}_*\|_2=1$ ו- $\gamma>0$ כך שלכל $i=1,\dots,m$ מתקיים,

$$y_i(\mathbf{w}_*^\top \mathbf{x}_i)\geq\gamma$$

נניח בנוסף כי לכל $i=1,\dots,m$ מתקיים $\|\mathbf{x}_i\|_2\leq R$.

אזי אלגוריתם הפרספטרון עושה לכל היותר $\frac{R^2}{\gamma^2}$ טעויות.

בשאלה זו נוכיח משפט זה בשלבים. תהי S קבוצת אימון ונניח את קיומם של w_*, R, γ כמו במשפט.

נגדיר את $\mathbf{w}_k$ להיות וקטור המשקולות שנלמד עד הטעות ה- k (לא כולל).

<u>שימו לב</u>: עבור הגדרה זו מתקיים $\mathbf{w}_1=\mathbf{0}_d$.

א. [2 נק'] האם S פרידה לינארית? נמקו בקצרה.

<u>פתרון</u>: כן. נתון כי קיימים w_* ו-$\gamma>0$ כך שמתקיים

$$y_i(\mathrm{w}_*^\top x_i)\geq\gamma\geq 0$$

זה אומר כי הסיווג של w_* לכל דוגמה x_i הוא נכון הרי $w_*^\top x_i$ וגם y_i עם אותו סימן.

<u>טעות נפוצה</u>: S לא פרידה לינארית כי לא נתון עליה כלום. אבל כן נתונים w_*, R, γ כמו במשפט.

ב. [5 נק׳] נתחיל בלהוכיח באינדוקציה את הטענה הבאה:

$$\mathbf{w}_{k+1}^\top \mathbf{w}_* \ge k\gamma$$

<u>הדרכה:</u>

- בדקו את בסיס האינדוקציה עבור $k = 0$.
- הניחו את נכונות הטענה עבור k כלשהו, כלומר: $\mathbf{w}_k^\top \mathbf{w}_* \ge (k-1)\gamma$.
- הוכיחו באמצעות כלל העדכון של הפרספטרון כי $\mathbf{w}_{k+1}^\top \mathbf{w}_* \ge \mathbf{w}_k^\top \mathbf{w}_* + \gamma$.
- השלימו את צעד האינדוקציה.

<u>פתרון:</u>

עבור $k = 0$ מתקיים $\mathrm{w}_1^\top w_* = 0^\top w_* = 0 \ge 0 \cdot \gamma = 0$.

נניח את הטענה עבור k כללי ונוכיח עבור $k + 1$.

$$\mathrm{w}_{k+1}^\top w_* \ge (w_k + y_i x_i)^\top w_* = w_k^\top w_* + y_i x_i^\top w_* \underbrace{\ge}_{\substack{y_i x_i^\top w_* \ge \gamma \\ w_k^\top w_* \ge (k-1)\gamma}} (k-1)\gamma + \gamma = k\gamma$$

צעד האינדוקציה הושלם.

ג. [2 נק׳] לרשותכם אי-שוויון קושי שוורץ

$$\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \ge |\mathbf{u}^\top \mathbf{v}| \qquad \mathbf{u} \in \mathbb{R}^d, \mathbf{v} \in \mathbb{R}^d$$

הסיקו

$$\|\mathbf{w}_{k+1}\|_2 \ge \mathbf{w}_{k+1}^T \mathbf{w}_*$$

<u>פתרון:</u> נבחין כי עד סעיף זה הוכחנו כי $\mathrm{w}_{k+1}^\top \mathrm{w}_* \ge k\gamma \ge 0$ ולכן $|\mathrm{w}_{k+1}^\top \mathrm{w}_*| = \mathrm{w}_{k+1}^\top \mathrm{w}_*$. מהאי-שוויון נסיק

$$\|\mathrm{w}_{k+1}\|_2 \ge |\mathrm{w}_{k+1}^\top \mathrm{w}_*| = \mathrm{w}_{k+1}^\top \mathrm{w}_*$$

<u>טעות נפוצה:</u> לא להתייחס לערך המוחלט.

סה״כ מהסעיפים הקודמים מתקבל $\|\mathbf{w}_{k+1}\|_2 \geq k\gamma$. כעת ניגש לחסם מלעיל של הביטוי.

ד. [9 נק׳] הוכיחו באינדוקציה את הטענה

$$\|\mathbf{w}_{k+1}\|_2^2 \leq kR^2$$

<u>הדרכה:</u>

- בדקו את בסיס האינדוקציה עבור $k = 0$.
- הניחו את נכונות הטענה עבור k כלשהו, כלומר: $\|\mathbf{w}_k\|_2^2 \leq (k-1)R^2$.
- הוכיחו באמצעות כלל העדכון של הפרספטרון כי $\|\mathbf{w}_{k+1}\|_2^2 \leq \|\mathbf{w}_k\|_2^2 + R^2$.
- השלימו את צעד האינדוקציה.

<u>זכרו:</u> עבור $\mathbf{v} \in \mathbb{R}^d$ מתקיים $\|\mathbf{v}\|_2^2 = \mathbf{v}^\top \mathbf{v}$.

<u>פתרון:</u>
עבור $k = 0$ מתקיים $\|w_1\|_2^2 = 0 \leq 0 \cdot R^2 = 0$.

נניח את הטענה עבור k כללי כלשהו ונוכיח עבור $k + 1$.

$$\|w_{k+1}\|_2^2 = \|w_k + y_i x_i\|_2^2 = (w_k + y_i x_i)^\top (w_k + y_i x_i) = w_k^\top w_k + y_i^2 x_i^\top x_i + 2y_i w_k^\top x_i =$$
$$= \|w_k\|_2^2 + \|x_i\|_2^2 + 2y_i w_k^\top x_i \underbrace{\leq}_{y_i w_k^\top x_i \leq 0} \|w_k\|_2^2 + \|x_i\|_2^2 \leq \|w_k\|_2^2 + R^2 \underbrace{\leq}_{induction} (k-1)R^2 + R^2 = kR^2$$

הסיבה לזה ש- $y_i w_k^\top x_i \leq 0$ היא ש- $w_k^\top$ טועה על הדוגמה ה- x_i.
צעד האינדוקציה הושלם.

<u>טעויות נפוצות:</u>

1. חלקכם פשוט החלטתם להעלים את הגורם $y_i w_k^\top x_i$ ללא נימוק, והוא ה-״קושי״ המרכזי בסעיף זה.
2. שימוש לא נכון באי-שוויון המשולש. טענתם

$$\|w_k + y_i x_i\|_2^2 \leq \|w_k\|_2^2 + \|y_i x_i\|_2^2$$

ניסוי מחשבתי קצר סותר טיעון זה, הרי אם הוא היה נכון אז

$$4 = (1+1)^2 \leq 1^2 + 1^2 = 2$$

סה״כ מכל הסעיפים עד כה קיבלנו

$$k^2\gamma^2 \leq \|\mathbf{w}_{k+1}\|_2^2 \leq kR^2$$

ה. [2 נק׳] הסיקו כעת את המשפט. כלומר:

$$k \leq \frac{R^2}{\gamma^2}$$

כאשר k הוא מספר הטעויות.

<u>פתרון:</u> אם $k = 0$, אזי המשפט מתקיים הרי $R \geq 0, \gamma > 0$. אחרת,

$$k^2\gamma^2 \leq kR^2 \Rightarrow k \leq \frac{R^2}{\gamma^2}$$

<u>טעות נפוצה:</u> לא להתייחס למקרה $k = 0$.

שאלה 4: Deep Learning [20 נק']

☐ לזכאים להערכה חלופית בלבד (כפי שהוגדרו באתר הקורס): סמנו את התיבה הזו אם ברצונכם **לדלג** על שאלה זו. המשקל של יתר השאלות יתפזר באופן יחסי על פני 100 נקודות. ניתן לדלג רק על שאלה אחת מתוך שאלות 3,4.

בשאלה זו נעבוד מעל מרחב דוגמאות $\mathcal{X} = \mathbb{R}$ ומרחב תיוגים $\mathcal{Y} = \{0,1\}$. בסעיפים בהם נדרש חישוב, עגלו את תשובותיכם לדיוק של עד שלוש ספרות אחרי הנקודה.

נתונה רשת נוירונים שעוצבה לפתירת בעיית סיווג בינארי המוגדרת ע״י הקשרים הבאים:

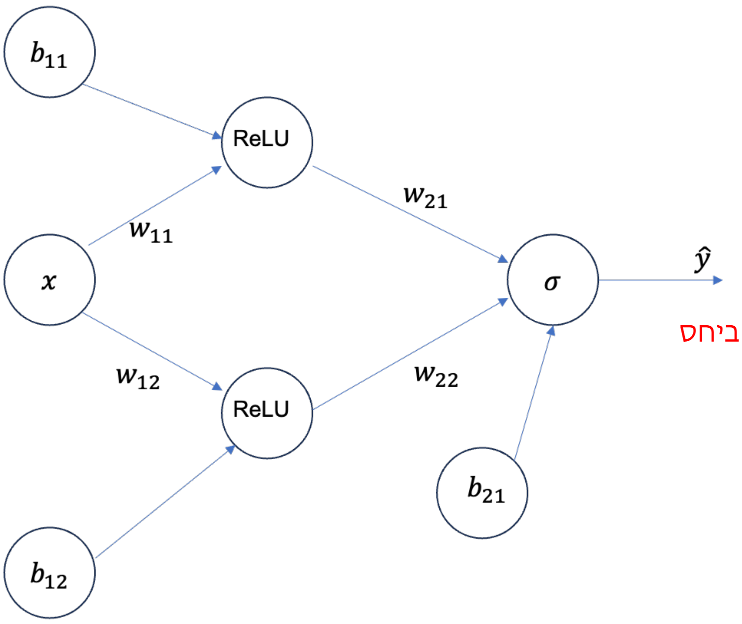
$$\begin{bmatrix} a_1 = & w_{11}x + b_{11} \\ a_2 = & w_{12}x + b_{12} \\ z_1 = & ReLU(a_1) \\ z_2 = & ReLU(a_2) \\ a_3 = & w_{21}z_1 + w_{22}z_2 + b_{21} \\ \hat{y} = & \sigma(a_3) \end{bmatrix}$$

כאשר $x \in \mathbb{R}$ הוא הקלט ו- $\hat{y} \in [0,1]$ הוא הפלט של הרשת.

האימון של הרשת מתבצע באמצעות cross-entropy loss. תזכורת:

$$\begin{bmatrix} ReLU(x) = & \max(0,x) \\ \sigma(x) = & \dfrac{1}{1+e^{-x}} \\ \ell^{CE}(y,\hat{y}) = & -ylog(\hat{y}) - (1-y)\log(1-\hat{y}) \end{bmatrix}$$

א. [2 נק'] ציירו תרשים של הרשת (לרשותכם דוגמה לתרשים בגוף התשובה).
ציינו מי הן המשקולות של הרשת שביחס אליהן אנו עושים אופטימיזציה.



אנו עושים אופטימיזציה ביחס ל:

$b_{11}, b_{12}, w_{11}, w_{12}, w_{21}, w_{22}, b_{21}$

טעות נפוצה:

ביקשנו שתציינו במפורש מי הן המשקולות של הרשת ביחס אליהן אנו עושים אופטימיזציה.

ב. [5 נק'] נניח כי

$$\begin{bmatrix} b_{11} = & 0.04 \\ b_{12} = & 0.01 \\ b_{21} = & 0.08 \\ w_{11} = & 0.2 \\ w_{12} = & -0.1 \\ w_{21} = & 0.7 \\ w_{22} = & 0.7 \end{bmatrix}$$

עבור קלט $x = 0.3$ חשבו את $\hat{y}$. אם ידוע כי הרשת מנבאת 1 אמ״מ $\hat{y} > \frac{1}{2}$, מה היא תנבא עבור $x = 0.3$?

פתרון: חישוב פשוט יניב כאן $\hat{y} \approx 0.53$, כלומר הרשת תנבא $+1$.

ג. [1 נק׳] באיזה כלל אנו משתמשים בשביל לחשב את הנגזרות החלקיות ביחס למשקולות של רשת נוירונים כלשהי?

סמנו את התשובה הנכונה:

1. כלל הפיצה.
2. כלל האצבע.
3. כלל השרשרת.
4. כלל השורש.

בסעיף הבא הניחו כי עבור הדוגמה $x = 0.3$ התיוג האמיתי שלה הוא $y = 1$.

נעסוק כעת באלגוריתם ה- backpropogation.

ד. [12 נק'] בצעו את אלגוריתם backpropogation על המשתנה b_{12}. עליכם לכתוב את הנגזרת החלקית של פונקציית ההפסד $\ell(y, \hat{y})$ ביחס למשתנה b_{12}, דהיינו $\frac{\partial \ell}{\partial b_{12}}$, באמצעות הנגזרות החלקיות $\frac{\partial \alpha}{\partial \beta}$, כאשר α, β יכולים להיות כל אחד מהבאים:

$$\ell, \hat{y}, z_i, a_i, b_{ij}, w_{ij}, x$$

עבור כל הערכים החוקיים של i, j. וודאו כי כל נגזרת חלקית $\frac{\partial \alpha}{\partial \beta}$ לא ניתנת לפירוק לנגזרות חלקיות פשוטות יותר.

לאחר מכן, **חשבו** את $\frac{\partial \ell}{\partial b_{12}}$ עבור $x = 0.3$. תזכורת: $\frac{d}{dx}\sigma(x) = \sigma(x) \cdot \left(1 - \sigma(x)\right)$

פתרון:

$$\frac{\partial \ell}{\partial b_{12}} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial a_3} \cdot \frac{\partial a_3}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_2} \cdot \frac{\partial a_2}{\partial b_{12}}$$

בזמן החישוב היה ניתן להבחין כי $\frac{\partial z_2}{a_2} = 0$, ולכן כל הנגזרת מתאפסת.